Arabic Information Retrieval System based on Morphological Analysis (AIRSMA):

A comparative study of word, stem, root and morpho-semantic methods

By

Musaid Saleh Al Tayyar

BA Lib. & Inf. Sci. Imam University
MA Lib & Inf. Sci. Loughborough University

THESIS

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science in the Department of the Computer and Information Science
DeMontfort University

July 2000

بسم الله الرحمن الرحيم

In the name of Allah the Most Gracious, the Most Merciful

Acknowledgement

First of all, all the praise and thanks be to Allah the Lord of the mankind and all that exists.

Secondly, my deep thanks goes to my supervisor Dr. Kamal Bechkoum I extremely grateful for his guidance and advice.

I should also thank my second supervisor Dr. Gordon Clapworthy for his support and suggestions.

I should also like to express my gratitude to my family and to my loyal wife and my children for their encouragement and for the understanding and patience they have shown during my period of study.

I would also like to extend my gratitude and appreciation to those people whom cooperation and help during the queries formalisation and the relevance judgment tasks.

I am most grateful to the Imam Mohammed Ibn Saud Islamic University for providing me the scholarship to pursue my studies at De Montfort University.

Dedication

This thesis is dedicated to my dear parents,
to my loyal wife and my children: Tamathor,
Naila and Arwa for without their support this
work would not have been possible

Abstract

Text retrieval systems in Arabic are new phenomena. A number of techniques such as truncation, stemming algorithms, and morphological analysers have been introduced into text retrieval systems to improve the retrieval performance. In Arabic information retrieval systems, three search methods are used namely: word, stem, and root. The word method is based only on term matching, while the other two methods use morphological analysis. The two methods have different levels of morphological analysis. However, each of these has its limitations. For example, the word and stem methods may miss relevant records (because of morphological variations). On the other hand, the root method may retrieve irrelevant records. This due to the fact that the root method is capable of reducing a given word to its root, and then generate all possible morphological variations of that word. The limitations of the current search methods have motivated this researcher to investigate a novel method to be used in an Arabic information retrieval system. This approach is called the morpho-semantic method. The morpho-semantic method is based on semantic links of the morphological forms. This approach uses a representative sample of the Arabic morphological forms (nouns and adjectives) which form the basis of the majority of Arabic words. The aim of introducing this method is based on the hope that this method will improve the effectiveness of the word and stem methods in terms of retrieving more relevant records than the word and stem methods did. At the same time, it is also hoped that the same method will improve the root method in terms of rejecting the irrelevant records that may be retrieved by the root method. A prototype text retrieval system was developed using Prolog. The system prototype can be used as a search engine for an Arabic text database or Internet sites. To investigate the retrieval performance of each, a sample of 590 records in Arabic was used as a database. Recall and precision are used to measure and evaluate the effectiveness of the word, stem, root and morpho-semantic methods. Regarding the performance of the recall parameter, it was found that the morpho-semantic method had the highest level with 91%, then the root method with 88%, followed by the stem method with 79%. The lowest recall was achieved by the word method at level 54%. On the other hand, the word method had the highest level of precision at 82%, then the stem method at 78%, followed by the morpho-semantic at 67%, and finally the root method with the lowest level of precision at 59%. As far as morpho-semantic is concerned, the recall level shows the superior nature of this method over other methods. However, the same method shows the superiority of the root method in terms of precision. The findings of the study indicate that the morpho-semantic method does improve the retrieval performance of the word and stem methods as far as recall (i.e. more relevant records retrieved) is concerned. Furthermore, the same method improves the retrieval performance of the root method in terms of precision (i.e. fewer irrelevant records retrieved).

Table of contents

Chapter 1: Introduction

Chapter 2: Information retrieval and language

Chapter 3: Arabic language stricture

Chapter 4: Literature Review

Chapter 5: Experimental design and methodology

Chapter 6: Morphological system in Arabic

Chapter 7: System Architecture

Chapter 8: System evaluation

Chapter 9: Summary and conclusion

List of Tables

List of Figures

Chapter One: Introduction

1.1 Introduction

The development of Arabic text retrieval systems is a relatively recent phenomenon. In the last decade the Arab world has witnessed a number of attempts to develop Arabic text retrieval systems. The current study is one of these attempts. However, since the computer was introduced to the Arabic text and information retrieval environment, a number of problems have arisen (for example language issues such as morphology). Some of these problems have been solved, while others remain unsolved. The main aim of this study is to participate in solving some of the linguistic issues, specifically those issues related to morphology. This chapter starts with a detailed description identifying the problem of the study. This is followed by the aims of the study. The chapter also highlights the significance of the study, in addition to its scope and limitation. A brief definition of the terms being used in this thesis is given. The chapter ends with chapter summary of the thesis.

1.2 General problem statement

The Arabic language belongs to the Semitic family of languages. The words in such languages may be formed by modifying the root itself internally and not simply by the concatenation of affixes and roots as occurs in an inflecting (such as Latin), agglutinating (such as Turkish), or incorporating language (Greenlandic Eskimo) (Katamba, 1993). This type of processing is known as morphology. Arabic morphology has a great impact on word formation. An Arabic word may appear in a text in different morphological variations. For example, a word such as كِتَاب (book) may appears in a text as كتابان (two books), كُتُب (books), كِتَابه (his book), كتابها (her book), كُتُبه (his books), and so forth.

Morphological variations are one of the many characteristics of natural language that must taken into account when designing a free-text retrieval system, since there may be some, or many, forms of a given word. These forms result from the addition of different prefixes and suffixes or even infixes to a common word stem according to the dictates of grammar as mentioned above (Popovic and Willett, 1992). For example, if the retrieval system is based on word search only then it will retrieve the word as it is entered. However, if the system is based on morphological analysis searching, the system will retrieve all morphological variations of the word كتاب (book) as mentioned above.

Using morphological analysis to support information retrieval in Arabic has lead to some differences of opinion between some computer specialists (Ali, 1988; Al Fedaghi and Al Anzi,1989) and some information specialists (Al Atram 1989; Al Swaynia 1994) as to the most suitable methods for information retrieval (word, stem or root methods). Some computer specialists believe that the Arabic information retrieval system should be based on morphological analysis in order to retrieve the word by its root. They justify their view by stating that Arabic is a derivative language, therefore the system should be based on the root of the word. They give the following example: suppose that the user wants to search for the word كتب "*kataba*" (wrote); if the system is based on word search only then it will retrieve the word as it is entered. However if the system was based on root searching, the system will retrieve all forms of the root "*ktb*" such كتب *kataba* (he wrote), يكتب *yaktubu* (he is writing), نكتب *naktubu* (we are writing), تكتب *taktubu* (she is writing) etc.

Some information specialists (Al Atram, 1989; Al Swaynia, 1994), on the other hand, do not believe there to be a useful relation between information retrieval and the root of the word. Among examples they give the following: from the root جمع "*jma*" we can derive a number of words: اجتماع *ijtima* (meeting), جمع *aljama* (summation),جمعة *aljumah* (Friday), جماعة *jama'ah* (group), جامع *aljamia* (mosque), جامعة *aljamiah* (university). Thus, they justify their rejection of the root method by stating that if the information retrieval were to be based on this system it would retrieve all the above words though there is no real need to do so.

Unfortunately, neither group offered clear evidence for their claims. As was mentioned above, the rejection or support for the root method was only based on giving examples. As far as academic research is concerned, it is not enough to say that the root method is function better or worse than the other methods without any proof to support or reject your belief. However, some academic work such as that of Al Kharashi (1991); Abu Salem (1992); and Hmeidi (1995) offer a number of experiments to investigate the retrieval performance of the following methods (word, stem, and root). Full details of these works can be found in Chapter 4.

In order to overcome the morphological variations of the word, a number of techniques such as truncation, stemming algorithms, and morphological analysers have been introduced into information retrieval systems to improve the retrieval performance. In Arabic information retrieval systems, three search methods are used: namely, word, stem, and root. The word method is based only on term matching, while the other two methods are based on morphological analysis. These have different levels of morphological analysis. However, each of these has its limitations. For example, the word and stem methods may miss relevant records (because of morphological variations). On the other hand, the root method may retrieve irrelevant records. This due to the fact that the root method is capable of reducing a given word to its root, and then it will generate all possible morphological variations of that word.

The limitations of the current search methods have motivated this researcher to investigate a novel approach to be used in an Arabic information retrieval system. This approach has been given the name " the morpho-semantic method". It is hoped that this method will improve the effectiveness of the word and stem methods in terms of retrieving more relevant records than the previous word and stem methods did. At the same time, it is also hoped that the same method will improve the root method in terms of rejecting irrelevant records that may be retrieved by the root method. This approach uses a representative sample of the Arabic morphological forms (nouns and adjectives) which form the basis of the majority of Arabic words. A detailed discussion of the morpho-semantic method is given in Chapter 7.

To address the problems identified above, a prototype information retrieval system was developed using Prolog. The prototype is based on four search methods: namely, word, stem, root, and morpho-semantic methods. The first methods (i.e. word, stem and root) have been used in some commercial software, in addition to a few academic studies, whilst the morpho-semantic method has not yet been used (as far as the writer of this thesis knows). Moreover, although the stem method is used in some commercial software, it has still not been developed to its potential. Further details about each method are discussed in Chapter 8. The aims of developing the prototype are discussed in the next section.

1.3 Aims of the study and research questions

The main aim of the study is to improve the retrieval performance in Arabic texts retrieval systems. In order to achieve this aim the following objectives were identified.

1- To investigate the retrieval performance of the following methods: word, stem, and root;

2- To investigate and implement a novel method of search (morpho-semantic) based on the semantic links between morphological forms and compare the retrieval performance of this approach with the above methods;

3- Designing a morphological analyser and generator to support information retrieval in Arabic;

4- To investigate the effectiveness of the morphological analysis (derivational and inflectional morphology) on information retrieval performance.

The study also aims to answer the following questions:

- Does the morpho-semantic method improve retrieval performance of word and stem methods in terms of recall?

- Does the morpho-semantic method improve retrieval performance of the root method in terms of precision?

- Do the morphological variations have an effect on retrieval performance in Arabic?

□ Why do some methods retrieve more relevant records than others, or vice versa? Why do some methods retrieve more irrelevant records than the other methods did?

1.4 Significance of the study

As was mentioned above, this study aims to develop a novel method of search to be used in an Arabic information retrieval system with a view to improve retrieval performance. This is the main significance and contribution of the study. Furthermore, so far very little research has been carried out to investigate the effectiveness of morphological analysis in information retrieval in Arabic. The study will draw attention to the importance of research into morphological analysis and its relationship to information retrieval in Arabic. It will also provide an open avenue for further work in Arabic language research as related to the development of text and information retrieval systems.

A further significance of the study is to design an Arabic information retrieval prototype based on morphological analysis. As was noted form reviewing the Arabic literature related to information retrieval systems, a few numbers of information retrieval systems are available to be used for Arabic document retrieval. The prototype being developed by the study can be promoted later on and used as a search engine for Arabic sites on the Internet.

Another significance of the study is the implementation of the morphological analysis for information retrieval in Arabic. The morphological analysis used in this study is based on two methods of morphology: derivation and inflection. Each of these methods has a specific impact on retrieval performance. The study deals with the two methods and shows the effectiveness of each one. Moreover, the study used some artificial intelligence techniques, such as semantic networks and rules, to represent Arabic morphological forms. The knowledge base of the morphological forms and the semantic links between them can be further developed and used, later on, as a basis for automatic indexing for Arabic texts.

Finally, the results of the study show how the morpho-semantic method has improved the retrieval performance of the root method in terms of precision (i.e. fewer irrelevant records retrieved). On the other hand, the same method has also improved the retrieval performance of the word and stem methods in terms of recall (i.e. more relevant records retrieved).

The superiority of the morpho-semantic method over the word and stem methods for recall was expected. What was not expected was the superiority of the morpho-semantic method over the root method. The recall of the morpho-semantic method was at a level of 91%, while the recall for the root method was at 88%. In theory, it was expected that the root method would retrieve more relevant records than the other methods (including the morpho-semantic method). This expectation was based on the fact that the root method is capable of retrieving most, if not all, of the morphological variations of a given word. It was found that the reason for such a performance for both methods was related to the Arabised word being used in a query or texts such as المجلات الإلكترونية (electronic journals). In other words, the morpho-semantic method was successful in retrieving them, while the root method was not. The performance of the morpho-semantic method shows us the significance and the potential use of this method when used as a search method for Arabic information retrieval systems.

1.5 Scope and limitation of the study

In any research or work there is a certain scope and some limitations; this study is no exception. First of all, the study concerns only Arabic information retrieval systems. Furthermore, the study evaluates the effectiveness of morphological analysis on retrieval performance, while other types of linguistic analysis, such as syntactic analysis, semantic analysis, and pragmatic analysis are beyond the scope of the study. Within the morphological analysis, the study does not cover all the morphological subjects. A representative sample of the morphological forms (nouns and adjectives) is used. This means that all the morphological forms of verbs, adverbs, and weak forms are not included in this study.

It is expensive to develop a comprehensive morphological analyser and generator for Arabic within the time being allowed to the current researcher. However, the morphological analyser and generator used in this study fulfill the task of running the experiments.

1.6 Definitions

The study uses specific terms related to language. It will be helpful if the reader is aware of the meaning of these terms. This section aims to give a brief definition of each term. Most of these terms were derived from Fromkin and Rodman (1998); Matthews (1997); and Katamba (1993).

- □ *Affixes*: bound morphemes attached to a stem or root morpheme (examples of affixes are prefix, infix, and suffix).

- □ *Derivation*: is a branch of morphology (derivational morphology), which is concerned with the derivation of one word in the lexicon from another: e.g. that of *keeper* from *keep*, or of *hopeless* from *hope*.

- □ *Inflection*: a branch of morphology (inflectional morphology), which is concerned with inflectional categories such as tense, voice, and number. The function of the inflectional categories may change the form of the word, but does not change the nature of it. Thus plural *books* is distinguished from singular *book* by the inflection *-s*, which is, by that token, a plural inflection.

- □ *Morpheme*: the smallest unit of linguistic meaning or function.

- □ *Morphology*: a branch of linguistic studies that deals with the internal structure of words.

- □ *Prefixes*: bound morphemes that occur before the root or stem of a word.

- □ *Root*: non affix lexical content morpheme, which cannot be analysed into smaller parts.

- □ *Stem*: a root morpheme combined with affix morphemes; other affixes can be added to a stem to form a more complex stem.

- □ *Suffixes*: bound morphemes that occur after the root or stem of a word.

1.7 Summary of the chapters

The present thesis is organised into nine chapters entitled: introduction; information retrieval and language; Arabic language structure; literature review; morphological system of Arabic; system architecture; system evaluation; and conclusion.

Chapter One of the thesis is mainly an introduction to the study which includes a problem statement and the aims of the study, in addition to the research questions, the significance of the study, the scope and limitation of the study, and finally a summary of the chapters.

Chapter Two deals with the background relating to the study. The background covers the linguistic issues which have an effect on information retrieval such as synonyms and morphological variations. It also represents the techniques used to improve retrieval performance such as truncation and Boolean operator.

Chapter Three describes the structure of the Arabic language; especially those elements which are related to information retrieval. The chapter gives an overview of Arabic.

Chapter Four covers the close-related literature. The chapter starts with a brief background of the information retrieval evaluation. It is then followed by the information retrieval systems in Arabic. The chapter discusses natural language processing (NLP) in Arabic. It is also contains a description of some morphological analysers for Arabic and presents a brief review of them. Chapter Four ends with a discussion of the relationship between morphological analysis and information retrieval.

Chapter Five is a detailed description of the experimental research design and methodology. The parameters used in this study were covered in the chapter. Four parameters are used to evaluate the retrieval performance of each method: namely, recall, precision, false drop, and omission factor are discussed in this chapter.

Chapter Six deals with the conceptual framework of the study. The chapter discusses in details two morphological theories (i.e. inflectional and derivational). The two theories were implemented by the current study.

Chapter Seven is a detailed description of the system architecture and implementation. The chapter describes the prototype and its components. It also covers how the inflectional and derivational morphology was implemented. The search methods of the system are discussed in this chapter. There are four search methods which have been evaluated by the current study: namely, word, stem, root, and morpho-semantic methods. The chapter ends with a description of the morphological analyser and generator which was developed to support information retrieval.

Chapter Eight covers the main system evaluation. An attempt was made to represent the retrieval performance of each method, in addition to offering a discussion of the results of each method. The retrieval performance evaluation is based on the four parameters which were mentioned in Chapter 5. The chapter also discusses the failure analysis for each method.

Chapter Nine is the last chapter of the thesis. It is a summary of the work which has been carried out in the current study. It also shows the main findings of the system evaluation and attempts to answer the research questions. The chapter presents several recommendations to those involved in information retrieval and natural language processing. The chapter ends with some suggestions for future work to be done in this area.

Chapter Two: Information retrieval and language

2.1 Introduction

In our daily life we need to store and retrieve information. This can be done either by using our mind (internal memory) or using external memory (e.g. machines or paper). Information retrieval is a challenge for those who are involved in this area. This challenge is due to the fact that a mass of data is produced every day and is available in an electronic format. These data are represented in different language styles. Covering all the areas related to this issue is beyond the scope of the chapter. The aim of this chapter is to make the reader aware of the relationship between the linguistic issues and information retrieval. It starts with a brief background of information growth, followed by an explanation of how information specialists and librarians deal with these complex issues.

2.2 Information growth

Information growth, or overload, has become one of the most significant problems facing information retrieval users whether they are using an online database or the World Wide Web. Information retrieval systems today are facing huge and heterogeneous masses of text data. The major problem of information retrieval is finding what a user wants (relevant documents) without wasting his/her time. With large volumes and masses of data, it is impossible for a human indexer to analyse document contents manually, therefore an automatic indexing method is needed in order to overcome some of the difficulties of information overload. A number of information retrieval tools have been developed to solve user queries or problems in finding relevant information in large collections of textual data. The following sections discuss most of them. Before this it might be useful to highlight some linguistic problems related to information retrieval.

2.3 Information retrieval and linguistic problems

As we know, human language consists of words, which come in sequences or sentences. With our words we express our ideas and we communicate our views. Because of our conversations the words of a language refer to things or have meaning. Some words are synonyms or mean the same; some are vague (Honderich, 1995). As a communication process, information retrieval relies on languages to carry out three major functions. These are as follows:

- First, languages are used to represent the content of documents, data, and other forms of information (Indexing).
- Second, the information problems of users are represented in terms of language (queries).
- Finally, languages are used to instruct the computer to carry out search and retrieval functions (Harter, 1986).

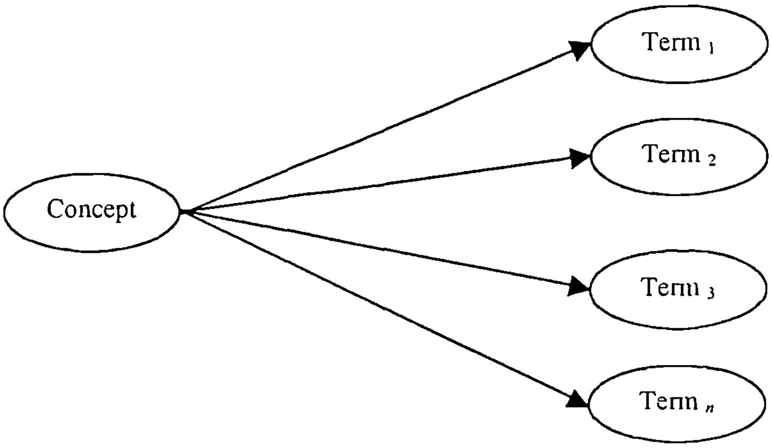
Using natural language for information retrieval has numerous problems. As Warner *el at* (1991) point out

> a major problem within information retrieval systems is referred to by Blair as the "indeterminacy" problem in subject access to documents. This indeterminacy arises because subjects are represented by various linguistic constructions; because different linguistic constructions can be used to represent the same meaning (a problem of synonyms); because the same linguistic construction may sometimes mean different things (a problem of ambiguity); and because there are other various and complex structural and semantic relationships among linguistic constructions. Indeed, one of the major problems facing those who undertake linguistic applications in retrieval has been to address the formidable computational problems in manipulating this very complex and poorly understood environment.

In human language (natural language) there are a number of ambiguities as mentioned above. The following sections discuss some of those which are most related to information retrieval.

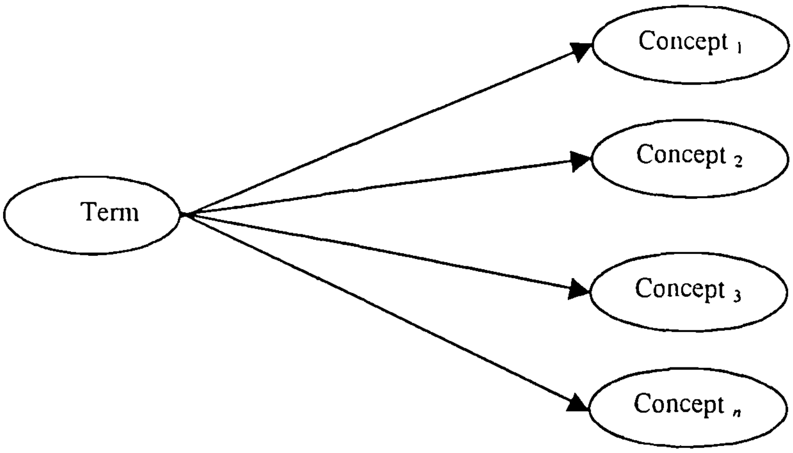
2.3.1 Synonyms

Synonyms are two or more different words that can be used to represent the same concept or meaning in the language. There are two types of synonyms: "absolute" synonyms which have meanings identical in all respects and in all contexts and "partial" (or near synonyms) which have meanings identical in some contexts e.g. in that replacing one with the other does not change the truth conditions of the sentence. Thus *paper* is a partial, though not an absolute, synonym of *article* (Matthews, 1997). This type of word in the language can be illustrated as below:



2.3.2 Homographs

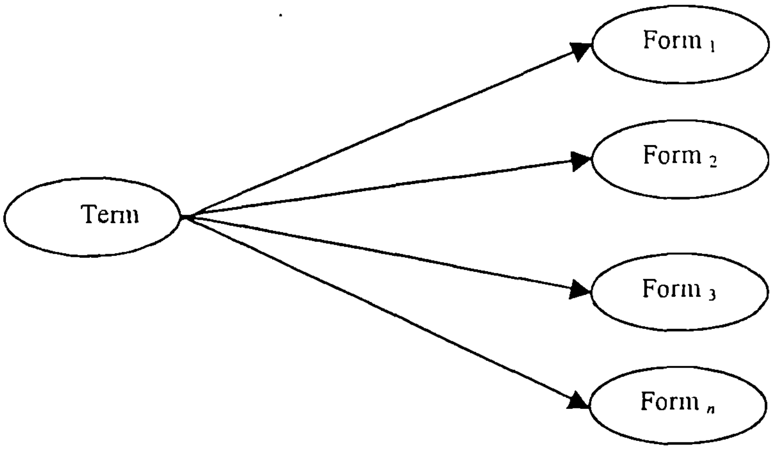
Homographs are the opposite of synonyms. They are different words spelled identically, and possibly pronounced the same (e.g. *lead* the metal and *lead*, what leaders do) (Fromkin and Rodman, 1998). They can be represented as follows:



As Harter (1986) mentions, if natural language is used as an information retrieval language, the existence of synonyms and near synonyms or homographs will increase the difficulty of information retrieval systems, because computers retrieve information by matching symbols, not the concepts represented by these symbols. Synonyms and homographs can be controlled and overcome by using controlled languages such subject headings and a thesaurus as an information retrieval language.

2.3.3 Word variations

This is another problem that arises as a result of using natural language as an information retrieval language. Word variation is a form of a specific word, either phonetic or orthographic, distinguished as such from the word as a lexical unit or lexeme: e.g. *ran* is one of a set of word forms (*run*, *runs*, *ran*, *running*) each of which is a form of the lexeme (to) *run* (Matthews, 1997). Word forms can be illustrated as shown below:



Word form (or morphological variation) is one of the many characteristics of natural language that must be taken into account when designing a free-text retrieval system. There are a number of techniques that have been used on information retrieval systems to overcome the problem of word forms. These techniques include suffix removal, truncation, stemming algorithms and morphological analysers.

Having identified the problems that occur when the natural language is used as an information retrieval searching language, and in order to overcome the above problems, information specialists developed a number of indexing languages such as thesaurus and subject headings to control synonyms or homographs and making concept relationships between terms. Other techniques such as morphological analysers, stemming algorithms and truncation are used to control the morphological variants of the words. The next section discusses the indexing languages which have been used by information retrieval systems to improve the retrieval performance of the systems.

2.4 Indexing languages

In the context of information retrieval, indexing can be defined as the process of allocating index terms, or keys, to a record or document. These index terms assist in the later retrieval of the document or record (Rowley, 1998). The assignment of indexing terms to the indexing language may be carried out manually (i.e. conducted by a human) or automatically (computer-based), or by a combination of the two approaches. Two types of indexing languages are used in information retrieval systems. The first type is known as *uncontrolled languages (or free text/natural language searching),* while the second type is known as *controlled languages (or vocabulary control).* The next section deals with both languages.

2.4.1 uncontrolled languages

If an information retrieval system is not based on indexes prepared manually, but on words and phrases as they occur within text itself, this type of system is known as free text search, natural language searching or uncontrolled language search. In such a system there is no human judgment involved in assigning specific terms to specific documents. The index terms are derived from texts directly. (See Figure 2.1.) The end user selects terms or a phrase or a group of words to describe his/her needs. As mentioned above, many linguistic problems such as synonyms, homographs, and word variations may occur when the uncontrolled language is used as an indexing language of the information retrieval system. A number of techniques were developed to enhance these systems. With automatic indexing (uncontrolled language) the following steps are followed (Salton and McGill, 1987):

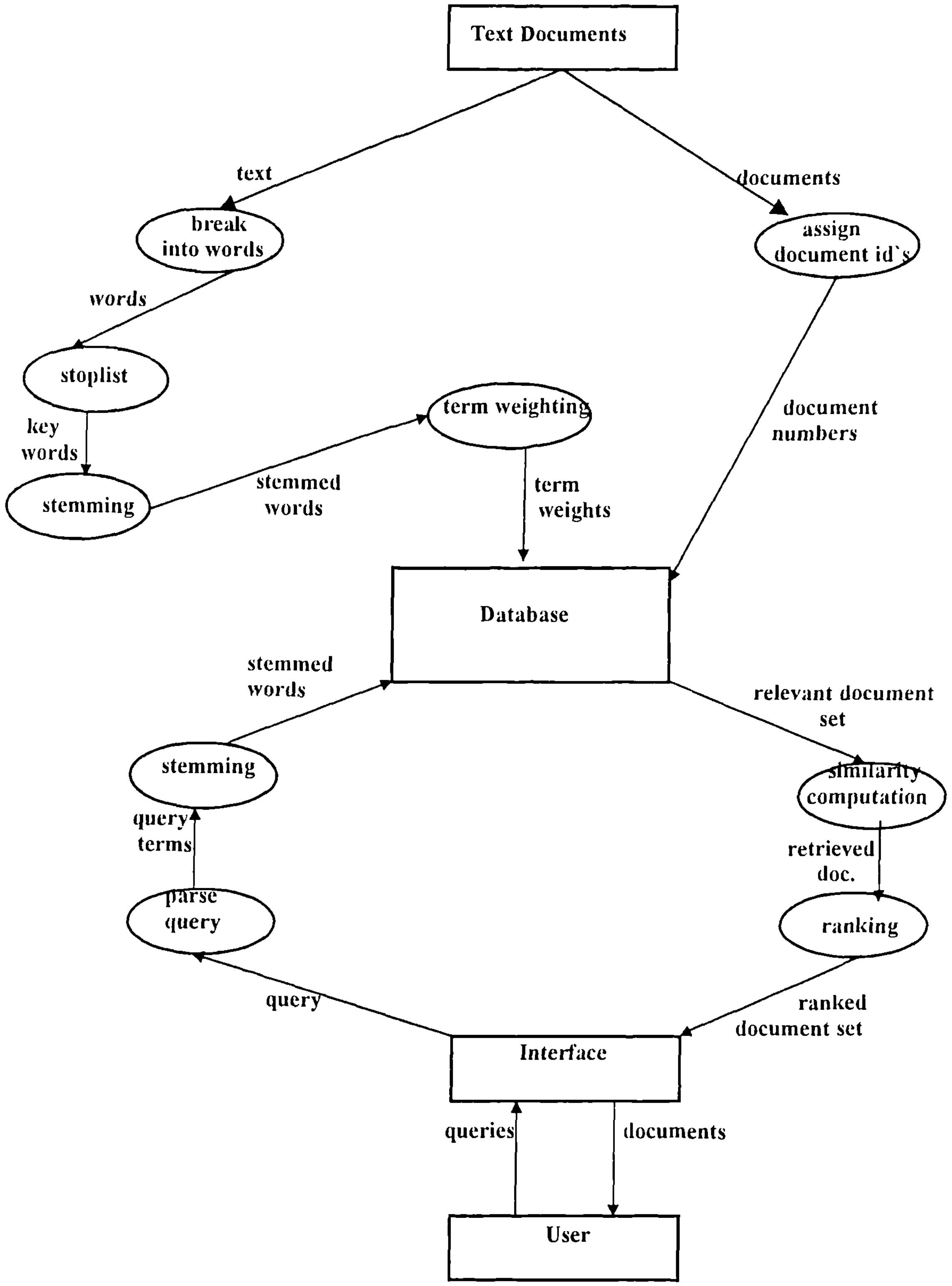


Figure 2.1 Processes involved in automatic indexing (adapted from Frakes and Baeza-Yates, 1992)

- all words of the text documents must be identified;
- the stop list words should be removed from the index files;
- the index terms must be identified and assigned to the documents of collection;
- all prefixes or suffixes that may be attached to index terms should be extracted (i.e. reducing the original words to word stem or root form);
- after the word stems or roots are generated, it becomes necessary to recognise equivalent stems occurring in the texts and to choose stems to be used as index terms (Searching point view). The above steps are depicted in Figure 2.1.

2.4.2 controlled language or vocabulary control

Controlled language is an authority list of terms linked to each other using cross-references. These lists can be used by the indexer to assign specific terms to specific documents based on subjective interpretations of the concepts in the documents. As Harter states (1986):

> There are several reasons for considering the use of an artificial language to represent information. There are many problems with use of natural languages for information retrieval, including synonymy, semantic ambiguity resulting from the existence of homographs, the semantic ambiguity inherent in 'soft' disciplines, false drops caused by contextual ambiguity, and the difficulty of performing generic searches. The use of controlled vocabularies can, to a large degree, solve these problems.

However, vocabulary control has been used in information retrieval systems to overcome the disadvantages of uncontrolled vocabulary. The vocabulary control can be broadly divided into two types: *subject headings*, and *thesaurus*. Using descriptors (as listed and described in a thesaurus) is another form of vocabulary control. Indexing via vocabulary control has three stages in a continuum:

- examining the documents and establishing subject content
- identifying the principal concepts in the documents
- expressing these concepts in the terms of the indexing language

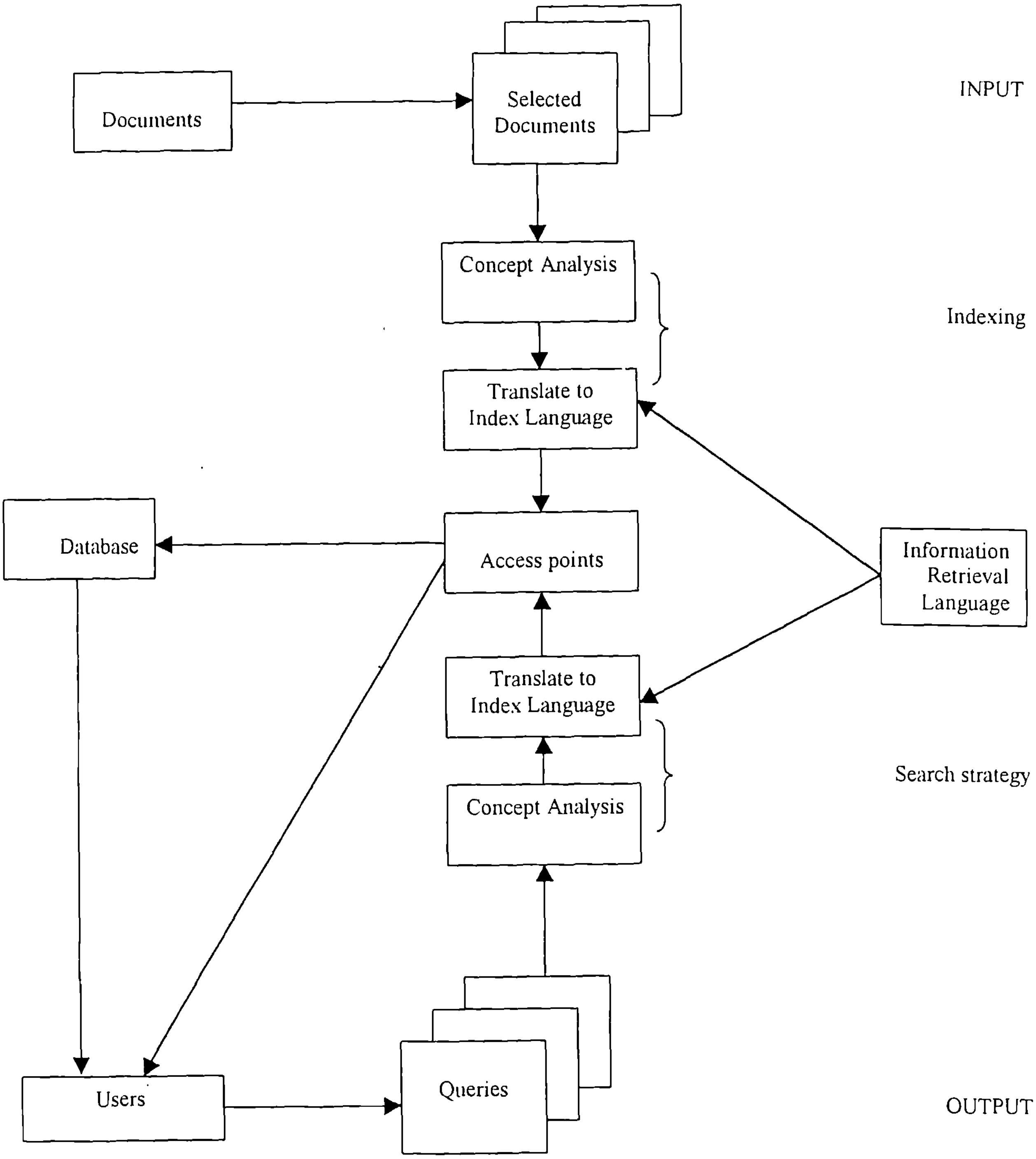


Figure 2.2 Processes involved in manual indexing (adapted from Lancaster and Warner, 1993)

As can be seen from Figure 2.2, in each stage there is a 'problem' requiring intellectual analysis. For conceptual indexing, the problem is a document or another piece of primary source material for which the major concepts are to be identified, from the point of view of the potential user population (indexing language). For information retrieval, the problem is to analyse the information needs of human users into its consistent concepts or facts, from the particular point of view of that user (searching language). In each case, the language used to express the summary statement is in natural language, the language of this information seeker or author. When the problem has been analysed into its constituent parts, a second step must be taken.

In conceptual indexing, the summary of the information item must be translated from natural language into the controlled vocabulary of the indexing language (Harter, 1986). To achieve a successful retrieval, the two elements (the access points of the documents and the user terms) must be closely related. Both indexing languages have their advantages and disadvantages. Table 2.1 summarises the advantages and disadvantages of both methods.

Table 2.1 Advantages and disadvantages of uncontrolled and controlled indexing languages (Rowley, 1998; Lancaster and Warner, 1993; Harter, 1986).

	Advantages	*Disadvantages*
Uncontrolled indexing language	Low cost; simplified searching; full database contents searchable; every word has equal retrieval value; no human indexing errors; no delay in incorporating	Greater burden on searcher: information implicitly but not overtly included in text may be missed; absence of specific to generic linkage; vocabulary of discipline must be known
Controlled vocabulary	Solves many semantic problems: permits generic relationships to be identified; maps areas of knowledge	High cost; possible inadequacies of coverage; human error, possibility of out-of-date vocabulary, difficulty of systematically incorporating all relevant relationships between terms; loss of precision; not specific

In addition to the above, several approaches such as statistical and linguistic analysis, have been developed during the last three decades to support information retrieval.

2.5 Statistical approach

The statistical approach is based on the assumption that the more frequently a word is used in a document, the more likely it is that the word is a significant indicator of the subject matter. The automatic selection of indexing terms is according to frequency of occurrence, according to prespecified rules (Feinberg, 1973). This approach was first suggested in the late 1950_s by Luhn (Luhn, 1957). Within this approach the computer is programmed to arrange the text words in increasing or decreasing order of frequency, and index terms for the documents are selected from this frequency list. Problems arise at this point because we know relatively little about the significance of high or low frequency. Because of the uncertainty concerning the relationship between the frequency of occurrence of a text word and its usefulness as an index term, subsequent research often combined frequency count with text reduction. Part of the text that is expected to contain useful information with a high degree of probability is subjected to a frequency count (Artandi, 1976). This approach is used to support the uncontrolled language system (the free text or natural language indexing).

2.6 Linguistic approach

Another approach to text analysis used for information retrieval is the linguistic approach. As Salton and McGill (1987) state, linguistic methods in information retrieval are really of two kinds: one the on hand, it is possible to use simple methodologies with limited aims, such as removing the ambiguity from some noun phrase identifier; on the other hand, more complex linguistic analysis systems can be utilized but the context in which these systems operate must be limited.

Natural Language Processing has been seen by a number of research groups (Arampatzis *et al.*, 1998) as a way to improve retrieval effectiveness. The simplest and most popular linguistic approach is stemming algorithms (i.e. removing prefixes or suffixes from a given word). In recent years, using natural language for information retrieval has become quite essential. Nevertheless, as Salton and McGill, (1987) acknowledge, the full scope of language understanding may not be needed in information retrieval.

2.7 Information retrieval techniques (enhancement IRS)

In an information retrieval environment it is well known that the relationship between a query and a document is determined primarily by the number and frequency of terms which they have in common. Unfortunately, words have many morphological variants, which will not be recognised by term-matching algorithms without some form of natural language processing. In most cases, these variants have similar semantic interpretations and can be treated as equivalent for information retrieval application (Hull, 1996).

In order to overcome the morphological variations of the word, a number of tools have been introduced to the information retrieval environment such as truncation, morphological analyser, prefix and suffix removal and Boolean operators.

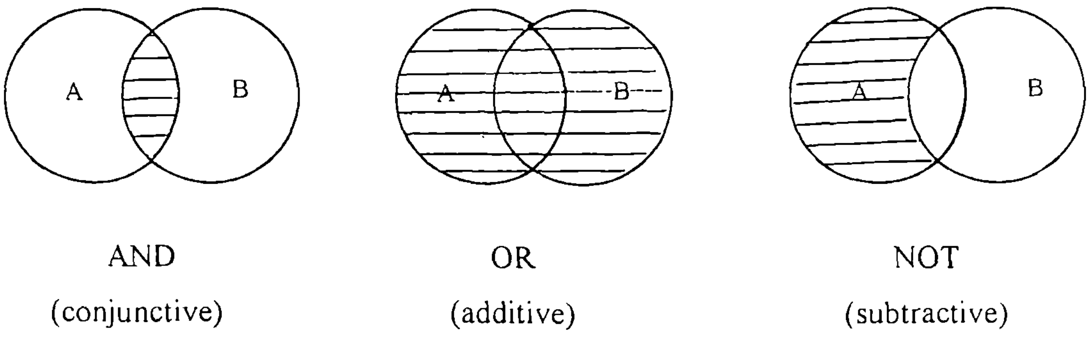
2.7.1 Truncation

Truncation supports searching on word stems. The use of truncation eliminates the need to specify each word variant, and thus simplifies the search strategies. The truncation technique is particularly useful in natural language information retrieval systems, where word variations are uncontrolled. There are two or three types of truncations: right-hand truncation (suffix removal), left-hand truncation (prefix removal) and middle truncation (infix removal). It is also useful for alternative spellings (Rowley, 1998).

The idea of truncation, as Paice (1996) states, is based on the fact that, for information retrieval purposes, what matters is the basic concept represented by the first few letters of a word; the ending represents the syntactic function or some other subsidiary property, and is a positive nuisance if it prevents corresponding ideas being matched to one another. Hence, if the endings can be removed or transformed, variant forms can be reduced to a common "stem" and thus treated as equivalent.

2.7.2 Boolean operators

Boolean operators (or search logic) are a very useful tool for information retrieval. They can be used to specify combinations of terms, which must be matched for successful retrieval. The Boolean operators can be used to solve problems of natural language such as synonyms, word variations, and spelling variations (English, American). The most well-known Boolean logic operators are AND, OR, and NOT. Venn diagrams (Figure 2.3) show the basic use of these operators.



AND: means both index tenns A and B must be assigned to a document for a match.
OR: means only one of the two index terms, A or B, need to be associated with a document for a match.
NOT: means the index term A must be assigned, and assigned in the absence of the term B for a match.

Figure 2.3 Boolean operators

2.7.3 Intelligent search (or Indexing) agents

The new technology used today to enhance information retrieval systems is *intelligent search agents* especially for the Internet search engines. As mentioned above, the amount of data which is available online is enormous. Agents have been developed as a solution to this problem. The intelligent search agents carry out a massive autonomous search of the Web. They can help in three ways (Finlay and Dix ,1996):

- ❑ They can find where suitable documents are stored.
- ❑ They can mediate between the user and different information sources.
- ❑ They can choose appropriate documents from large documents.

As Haverkamp and Gauch (1998) pointed out regarding information flow and the goals of intelligent search agents:

> Technology influences the amount and type of information available, but it must also provide the means to make effective use of this information from users' homes and desks. The research community could make a significant contribution by developing systems, which allow an end-user to search effectively. That is the

> goal of intelligent search agents, whether they search a single database of bibliographic records or a network of distributed, heterogeneous, hypertext documents.

2.8 Summary

Information retrieval systems today are facing huge and heterogeneous masses of text data As a communication process, information retrieval relies on languages to carry out three major functions: representing the content of documents, representing the information problems of users (queries), and retrieving functions. To carry out the retrieval procedure, two types of indexing languages are used in information retrieval systems. The first type is known as *uncontrolled languages (or free text/natural language searching),* while the second type is known as *controlled languages (or vocabulary control).*

For information retrieval, the problem is to analyse the information needs of human users into its consistent concepts or facts, from the particular point of view of that user (searching language). In conceptual indexing, the summary of the information item must be translated from natural language into the controlled vocabulary of the indexing language (Harter, 1986). Another approach to text analysis used for information retrieval is the linguistic approach. In recent years, using natural language for information retrieval has become quite essential.

Unfortunately, words have many morphological variants, which will not be recognised by term-matching algorithms without some form of natural language processing. To overcome these obstacles, a number of tools, such as truncation, Boolean operators and recently intelligent search agents have been introduced to support information retrieval systems.

Chapter Three: Arabic language structure

3.1 Introduction

The Arabic language shares, with the other natural languages, some common elements such as those mentioned in the previous chapter (i.e. synonyms, homographs, and word variations). The Arabic language is a member of the Semitic family of languages (different from Indo-European languages in some respects). It is spoken by over 150 million people in 21 Arab countries as the first language. An uncertain further number use it as a second language, mainly in Islamic countries. This chapter describes the structure of the Arabic language, although more emphasis is placed on those aspects which are important to information retrieval systems, such as affixation, broken plurals, morphology and derivation.

3.2 The alphabet

The Arabic alphabet consists of 28 characters which are called حروف الهجاء *hurof alheaja* (Table 3.1). From the twenty-eight characters there are three characters which appear in different shapes. These are as follows:

- *Hamza* [ء] is (sometimes) written with أ *alif* /a/ as in أكل *akala* (he eats), sometimes with ئ *ya* /y/ as in برئ *barea* (innocent), sometimes with ؤ *waw* /w/ as in سؤال *suaal* (question); or without any other characters as in قراءة *geraah* (reading) and similarly ملائم *mulaeem* (suitable).
- *Ta marbuta* [ة] the character [ه] with two dots above it is pronounced like letter */t/* in English. It is found only at the end of the word (nouns and adjectives) for example سنة *sanatun (*year).
- *alif magsurah* [ى] is the character ي *ya* /y/ without the dots below it. It is represented by the long vowel romanized as in مصطفى *mustafa* (masculine name).

The above three characters pose some difficulties in the setting up of an information retrieval system. Therefore some libraries and information centres ignore the *hamaz* and the two dots above *alta almarbutah* to unite the input and output for these characters. For example, consider a title such as: التنميـــة الاجتماعيـــة *altanmiatul alejtemaih* (social development). The indexer entered this title into the computer without the two dots above it. When users want to retrieve this title they must ignore the two dots otherwise the title will not be retrieved. Aman (1984) pointed out that, while the letters of the Latin alphabet have only one form, the sole exception being the use of capitals, this is not the case with Arabic script, where some characters appear in four, or possibly more different shapes. (See Table 3.1.)

3.3 The Arabic writing system

The Arabic writing system goes from right to left and most letters in Arabic words are joined together. Twenty-two among the twenty-eight can be joined on both sides and in the process take different shapes depending on their context in a word. The position can be at the beginning (initial) or in the middle (medial) or at the end (final) of the word (as shown in Table 3.1). The letter can also be written separately, not connected to another letter in the same word (isolated form).

3.4 Diacritical marks (vowelisation)

There is in Arabic a whole series of non-alphabetic signs, added above or below the consonant letters to make the reading of the word less ambiguous or absolutely certain. It must be emphasised that writers of Arabic do not normally use these signs except in very special cases. For example, the Quran is always fully "signed" to avoid any misreading. The same is often true of poetry, sometimes of foreign or unfamiliar words, and of beginners' manuals. The majority of these non-alphabetical signs relate to vowels (Wickens, 1980).

Table 3.1: Arabic Alphabet

Phonemic symbols for Arabic	Isolated Arabic Letters	Final Arabic Letters	Medial Arabic Letters	Initial Arabic Letters
A	ا	ـأ ، ء	ـئـ ، ء	أ
B	ب	ـب	ـبـ	بـ
T	ت	ـت	ـتـ	تـ
Th	ث	ـث	ـثـ	ثـ
J	ج	ـج	ـجـ	جـ
Ha	ح	ـح	ـحـ	حـ
Kh	خ	ـخ	ـخـ	خـ
D	د	ـد	——	د
Th	ذ	ـذ	——	ذ
R	ر	ـر	——	ر
Z	ز	ـز	——	ز
S	س	ـس	ـسـ	سـ
Sh	ش	ـش	ـشـ	شـ
S	ص	ـص	ـصـ	صـ
Tha	ض	ـض	ـضـ	ضـ
Ta	ط	ـط	ـطـ	طـ
Tha	ظ	ـظ	ـظـ	ظـ
a'a	ع	ـع	ـعـ	عـ
gh	غ	ـغ	ـغـ	غـ
f	ف	ـف	ـفـ	فـ
q	ق	ـق	ـقـ	قـ
k	ك	ـك	ـكـ	كـ
l	ل	ـل	ـلـ	لـ
m	م	ـم	ـمـ	مـ
n	ن	ـن	ـنـ	نـ
h	ه	ـه	ـهـ	هـ
w	و	ـو	——	و
y	ي	ـي	——	يـ

Although the majority of written Arabic texts are non-vowelised, the importance of vowelisation cannot be ignored. Vowelisation, in many cases, is necessary for resolving ambiguity in the meaning of some words. In fact, diacritical marks are essential for removing morphological ambiguity. For example, the word ذهب *dhb*, (to go) in the absence of vowelisation can be read as:

1- ذَهَبَ *dahaba* (to go)

or

2- ذَهَبٌ *dahab* (gold)

or

3- ذَهَّبَ *dahhaba* (to gild)

In some cases ambiguity can be removed through consideration of the context of a word in a sentence. In Arabic there are six vowels (Nasr, 1967); three are long ones, which appear in the alphabet; and three short ones, which do not appear in the alphabet but are added above or below the consonant letters.

3.4.1 Long vowels

These vowels appear as letters in a word and they are:

- □ (ا) /aa/ which is pronounced like the *a* in cat. For example مال (money).
- □ (و) /uu/ which is pronounced like the *oo* in pool. For example فول (bean).
- □ (ي) /ii/ which pronounced like the *ee* in meet. For example فيل (elephant).

3.4. 2 The short vowels

- □ فتحة (َ), above a consonant, pronounced like the *a* in cat (only much shorter). For example شك (doubt).
- □ ضمة (ُ), above a consonant, pronounced like the *u* in put. For example كن (be).

- □ كسرة (ِ), below a consonant, pronounced like the *i* in bit. For example طب (medicine).
- □ سكون (ْ), above a consonant pronounced like the *o* in factory, which indicates that the consonant is not followed by a vowel. For example وقف (entailing).

3.5 Affixation

According to *The Oxford Concise Dictionary of Linguistics* (Matthews, 1997), an affix is:

> Any element in the morphological structure of a word other than a root. E.g. *unkinder* consists of the root "kind" plus the affixes *un-* and *-er*. Hence affixation, for the process of adding an affix... Affixes are traditionally divided into prefixes, which come before the form to which they are joined; suffixes, which come after; and infixes, which are inserted within it.

Most Arabic words contain some kind of affixation. There is a relationship between affixation and morphology and derivation. (See Chapter 6.) Most Arabic words have either a prefix, a suffix or an infix. Sometimes all these affixations can be found in one word as in the word المزارعــون *almoZaReAun* (the farmers). The main function of morphological processors in Arabic is "to segment words into their individual morphemes, prefixes, infixes and suffixes in order to have real insight into the language. Each inflection holds important grammatical information such as number, gender, tense, definiteness, possessiveness and so forth" (Feddag, 1992). For example, the above-mentioned word can be decomposed into the following morphemes:

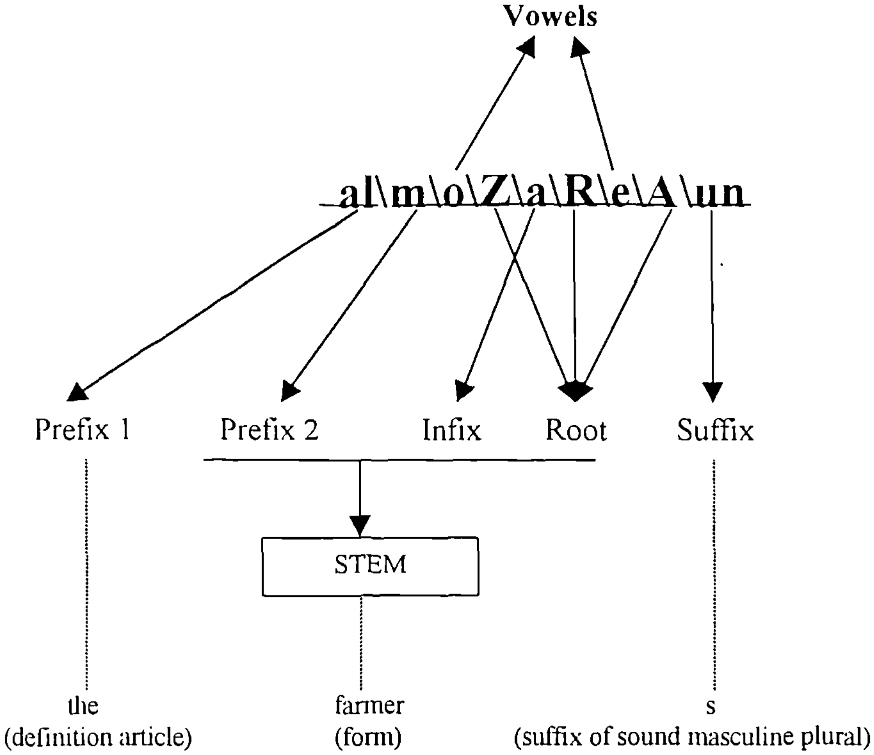


Figure 3.1 Word segmentation

3.6 The Arabic word classification

The study of the Arabic word is divided into two parts: the inflectional endings (*iarab*) and the changes that take place inside the word. The study of the first falls within the domain of syntax (*nahw*), and the latter falls within the domain of morphology. Arabic grammarians have given about equal treatment to syntax and morphology (the two are closely related in Arabic theory) (Owens, 1988).

The grouping of words into classes is highly dependant on the purpose of the classification. In Arabic, as in many languages, there are several methods for such classification. These include:

- ❑ A set of classes based on meaning, as in the traditional definition (noun: names, persons, places, things);
- ❑ A set of criteria based on the kind of endings words will take. For example, that any word having "*ing*" as a suffix is a verb;

- □ Another method of classification is based on the set of patterns in which a word can appear (noun, verb, and particle).

From the morphological point of view, an Arabic word can be classified into two parts (Al Hamlawi, 1991):

- □ Declinable nouns (meaning that the word can take three cases according to its position in the sentence),
- □ Plastic verbs (derived verbs).

For obvious reasons, morphological studies do not deal with the following word types (Al Aastee, 1992):

- □ Indeclinable nouns (meaning that the word has a fixed case, regardless of its position in the sentence)
- □ Aplastic verbs (non derived verbs), and
- □ Particles.

In summary, Arabic morphology treats only declinable nouns and plastic verbs. Nouns and verbs can be derived from the bare roots of the words by adding a set of letters to the roots (prefixes, infixes, and suffixes) or by changing the vowels of the root to generate words that are usually found in written and spoken Arabic. (See Chapter 6.) A word, from a morphological point of view, contains the following elements:

- □ Radical letters (a sequence of three or four valid characters from the alphabet);
- □ Augmented letters (10 letters from the alphabet which are added to the root to generate new words); known in Arabic as حروف الزيـادة (سالتمونيها) , and
- □ Vowels.

The combination of the above elements produces a word, either noun or verb, which may be segmented into the following structural features:

prefixes | stem (root + morphological form) | suffixes

3.6.1 Arabic words

As was mentioned above, Arabic words can be classified into different classes. The simplest method is dividing a word into three main categories: إسم *ism (*noun), فعل *feal* (verb), and حرف *harf* (particle). Each category can be subdivided into many types. A single word in Arabic could be a complete sentence, for example قامت *qamat* (she stood up), or even the same verb without the pronoun (ت) gives قام *qama* (he stood up).

3.6.1.1 Noun الإسم *al ism*

A noun in Arabic may be classified according to: number (singular, dual and plural), case (nominative, genitive and accusative), and definiteness/indefiniteness. The Arabic noun may be attached with three prefixes and three suffixes as shown below:

Prefixes	Stem	Suffixes
Pr. 1 + pr. 2 + pr. 3	Root + morphological form	su. 1 + su. 2 + su. 3

Prefix 1: elements serve only like conjunctions.

Prefix 2: elements and their associated elements determine the case of the noun.

Prefix 3: elements are used to indicate whether the derivational noun is declared with.

Suffix 1: elements determine whether the noun is feminine.

Suffix 2: elements and their associated attributes indicate the case of the noun and whether it is dual, masculine sound plural, or feminine sound plural and its case.

Suffix 3: elements are the pronouns attached to the derivational nouns. The elements and their features are the same as those of the object pronouns.

3.6.1.2 Verb الفعل *al fiil*

Arabic verbs have three main tenses: ماضى the past tense which is used for all actions which are already completed (e.g. كَتَبَ *kataba* [he wrote]), مضارع the present tense for all actions not yet complete (e.g. يكتب *yaktubu* [he writes]), and أمر the command form (e.g. أكتب *uktub* [do write]). Most Arabic verbs can be reduced to a past stem and a present stem, and a standard set of prefixes and suffixes can be added to these stems. The Arabic verb may be attached with two elements as prefixes and three elements as suffixes (an example is: سيتذكرونكم which means: (*they shall remember you*) each of which has a function (El Sadany and Hashish, 1989) as shown below:

Prefixes	Stem	Suffixes
Pr. 1 + pr. 2	root + morphological form	su. 1 + su. 2 + su. 3

Prefix 1: the elements of this list serve as conjunctions. Those associated with each element are used to indicate the tense and the case of the verb attached to it.

Prefix 2: the attributes associated with the elements of this list determine the tense of the verb and the features of the subject. The subject features are the person, gender, and number.

Suffix 1: the elements of this list are subject pronouns attached to the verb. The attributes associated with this list determine the tense, the case of the verb, and the subject features.

Suffix 2: the elements of this list are the first object pronouns. The associated elements here determine the features of the object pronoun (person, gender, and number) or the case of the verb.

Suffix 3: the elements of this suffix are the same as those in the suffix 2 list. In this case, they represent the second object pronoun or the case of the verb.

3.6.1.3 Particle الحرف *alharf*

The particle in Arabic is called حرف *harf,* meaning a letter. It is defined according to its functional category such as preposition, conjunction. For the purposes of work in information retrieval, most of the particles in the Arabic language would be on a stop list, so there is no need to discuss them in detail. However, some Arabic particles are joined to other words, such as ب *ba* (meaning: in) as بالمدينة *belmadine* (in the city) or ل *la* (meaning: for) للجامعات *lelljameat* (for universities). This joining creates problems for information retrieval in Arabic.

3.7 The number

According to Crystal's definition, a number is "a grammatical category used for the analysis of word classes, especially nouns which display such contrasts as singular, dual and plural" (Crystal, 1991a). All the above categories of numbers are used in Arabic.

3.7.1 The singular

The singular in Arabic is divided according to gender:

- □ A noun that refers to a male is masculine, such as مُدرِّس *mudarres* (teacher).
- □ A noun that refers to a female is feminine, for example مُدرِّسة *mudarresah* (teacher).

3.7.2 The dual

Arabic is one of the very few living languages which still has التثنية *altatheriah* (the dual) as a separate form which denotes there are two of something. The dual is formed regularly by adding the suffix ان */an/* (mark of the dual) in the nominative case, and يـن *in/* (mark of the dual) in oblique and accusative cases. For example: with قلم *qalam* (a pen) the dual is expressed by adding the suffix ان */an* to the singular, thus قلمـان *qalaman* (two pens) as this is in the nominative case. But in the oblique and accusative cases the suffix ين */in/* should be added to the singular thus: قلمين *qalamin* (two pens).

As seen above, the dual has changed according to its syntax case. A more detailed discussion of syntax in the Arabic language and its specific effects on sentences appears in Chapter 6. This type of formation has an effect on information retrieval.

3.7.3 The plural

In Arabic, there are two kinds of plural which are generally known as the sound (or strong) and the broken (or weak).

3.7.3.1 The sound plural

The sound plural is subdivided into two categories according to the gender as follows:

- ❑ The masculine sound plural which is formed by adding the suffix ون /un/ (mark of the masculine) to the singular in the nominative case. In the oblique and accusative cases the suffix ين /in is added to the singular.
- ❑ The feminine sound plural which is formed by adding the suffix ات /at/ (the mark of the feminine) in all syntax cases (nominative, oblique and accusative). For example, Table 3.2 shows some forms of sound plural.

Table 3.2 The Sound Plural suffixes

	Singular	Masculine		Feminine	
		Nominative	Oblique and Accusative	Nominative	Oblique and Accusative
	مدرّس	مدرّسون	مدرّسين	مدرّسات	مدرّسات
Arabic Translit.	mudarres	mudrres (un)	mudrerres(in)	mudrresatu	mudrresati
Translation	teacher	teacher(s)	teacher(s)	teacher(s)	teacher(s)

From the above table, it can be noted that in English 's' is added to make the plural, but Arabic is different. It depends upon the type of plural and whether it is masculine or feminine. It also has different shapes, which are called broken plurals, as will be shown in the next section. This changing of the suffixes causes some difficulty when the above words are to be retrieved.

3.7.3.2 The broken plural

This is the second type of plural in Arabic; it is also known as a weak or irregular plural. It was mentioned above that the sound plural is formed by suffixing ون or ين and ات depending on the type of plural, whether it is masculine (nominative or oblique) or feminine However, in the case of the broken plural, the matter is different because the broken plural has a number of measures which in Arabic is called أوزان *awzan*. Unfortunately, the singular word does not help in knowing them. The standard forms of broken plural number approximately 35 forms or measures (AbdAla'al, 1977).

There are more types of broken plurals, but these are not in common use. Also, some nouns have two or more different forms of broken plural such as بحر *bahar* (sea) which can be pluralised as بحور *buhur* or بحار *behar* or أبحر *abhur*. Hence, the broken plural in Arabic can cause some difficulty in information retrieval, particularly when natural language is used as an indexing language.

3.8 Morphology and derivation

Morphology has a great impact on word formation in Arabic. This, in turn, has an effect on information retrieval. As shown in Chapter 1, using morphology as a base for an Arabic information retrieval system is a debatable topic. Further details about the morphological system can be found in Chapter 6 of this thesis.

3.9 Summary

The Arabic language shares, with the other natural languages, some common elements such as synonyms, homographs, and word variations. The Arabic language is a member of the Semitic family of languages (different from Indo-European languages in some respects). The Arabic alphabet consists of 28 characters which are called حروف الهجاء *hurof alheaja*. The Arabic writing system goes from right to left and most letters in Arabic words are joined together. Most Arabic words contain some kind of affixation (either a prefix, a suffix or an infix).

The study of the Arabic word is divided into two parts: the inflectional endings (*iarab*) and the changes that take place inside the word. From the morphological point of view, an Arabic word can be classified into two parts: declinable nouns (meaning that the word can take three cases according to its position in the sentence), and Plastic verbs (derived verbs) Another type of classification is dividing the Arabic word into three main categories: إسم *ism (*noun), فعـل *feal* (verb), and حرف *harf* (particle). The noun and verb in Arabic may be further divided according to: number (singular, dual and plural), and case (nominative, genitive and accusative).

Chapter Four: Literature Review

4.1 Introduction

Since the computer was introduced to the Arabic information retrieval environment, a number of problems have arisen. Some of these problems have been solved, while others remain unsolved. In recent Arabic information retrieval literature an argument between some computer scientists and some information specialists has arisen about the effectiveness of using morphological analysis-based information retrieval. A number of morphological analysis algorithms have been reported in the literature. A limited number are used for information retrieval purposes. The following pages discuss the current state of these issues.

4.2 Information retrieval background and development

The goal of an information retrieval system is to search an information repository and retrieve records that are potentially relevant to a query. As Tague-Sutcliffe (1996a) stated: achieving this goal has always been a core activity of information professionals. Research into the best way to carry out this activity has been a concern of information science since the field arose in the $1950_{s.}$

Information retrieval experimentation is usually considered to have begun in the late 1950_s with the Cranfield tests by Cyril Cleverdon (Cleverdon,1966). In Cranfield I (1966), a comparative evaluation was made of four systems: the universal decimal classification (UDC), a conventional alphabetical subject index, a facet classification, and the Uniterm (keyword) coordinate indexing. The results showed all four systems retrieved the required paper in 74% to 88% of the cases, with the Uniterm system scoring the highest and faceted classification the lowest. Most failures were due to human error rather than the indexing system. Cranfield II, also carried out by Cleverdon, sought to determine the effect of specific recall-increasing and precision-increasing devices in a laboratory setting. The indexing languages evaluated consisted of single terms selected from the natural language of a document, alone, or combined with various devices such as synonyms, word form variants, quasi-synonyms, term weighting, links and roles, hierarchical linking of terms, phrases, and controlled language terms. The results showed that indexing, based on single natural language terms, performed best by this criterion.

Later experimenters, such as Salton (1971), have developed and used an experimental system known as SMART to investigate a number of sophisticated approaches. Among them are the word stem method, thesaurus, phrase etc. Early results showed that weighted word stem indexing, where the stems are extracted from the document texts, was more effective than such sophisticated methods as phrases and classification hierarchies.

For an excellent literature reviews on IR experimentation the reader is referred to Sparck Jones (1981); Sparck Jones (2000); Salton (1971); Salton and McGill (1987); Lancaster (1968); and Tague-Sutcliffe (1996a), in addition to several technical reports and articles published in periodicals and journals.

Information retrieval performance is influenced by a number of factors such as indexing language, search strategies, searchers etc. Fidel and Soergel (1983) discussed in detail most factors which is affect online bibliographic retrieval. They present a conceptual framework for the organisation of factors affecting the search, while Smithson (1994) in his paper outlines the problems involved in IR evaluation and argues for a more user-centred interpretative approach.

Language is among the factors which has an effect on information retrieval. Several authors, Salton and McGill (1987); Sparck Jones and Kay (1973); Perez-Carballo and Strzalkowski (2000) discussed linguistic issues related to information retrieval. Some authors, (Salton and McGill, 1987); (Doszkocs, 1986) believe that the full scope of language understanding may not be needed in information retrieval to achieve an acceptable level of performance. On the one hand, some authors have noted that grouping words which have the same root under the same stem increases the success of matching of documents to a query (Savoy, 1993; Harman, 1991; Porter, 1980; Montgomery, 1972).

Another major problem with IR is information growth or overload, which has become one of the most significant problems facing information retrieval users whether using online databases or the World Wide Web. Information retrieval systems today are facing huge and inhomogeneous masses of text data. To overcome this challenge, a number of tools have been developed and used. One of the most promising techniques is the use of the Intelligent Information Agents to support Internet search engines. Hundreds of articles, conference papers and web sites have been published considering the usage of the Intelligent Agents as information finders. For example, Haverkamp and Gauch (1998) provide an overview of and challenges for using intelligent information agents for distributed information sources.

Yang *et al* (1997) developed a natural language processing prototype based on Agent System (NIAGENT). The prototype presents users with a intelligent and friendly interface that understands a user's natural query, translates the user's interests into appropriate queries for each search engine, analyses the returned references and returns them. In order to compare the retrieval performance of the NIAGENT with other Internet (one example is MetaCrawler) search engines in terms of precision, the developers conducted several experiments. The results show that the NIAGENT achieved high precision at the level 0.90, while the MetaCrawler precision achievement was at the level 0.37. As the authors comment, the result is not surprising since NIAGENT takes one more step in analysing the contents of the reference web pages rather than using simple keyword search.

As is known, search engines are essential for finding information on the Internet. They provide three chief facilities (Gordon and Pathak, 1999). These are as follows:

- they gather together a set of Web pages that form the universe from which a searcher can retrieve information;
- they represent the pages in this universe in a fashion that attempts to capture their content;
- they allow searchers to issue queries, and they employ information retrieval algorithms that attempt to find for them the most relevant pages from this universe.

In the above paragraphs, an attempt was made to familiarise the readers with state-of-the art developments in general. The next sections will review most of the Arabic studies which were considered representative and closely related to the approach suggested in this study. The main concern here is to give more details about information retrieval in Arabic.

4.3 Information retrieval systems in Arabic

There are two approaches to the design of an Arabic information retrieval system. The first approach is to Arabize an existing system from another language (usually English) into a format which is capable of handling Arabic text. These will be called Arabised systems. The second approach is to build an Arabic system to handle Arabic text (i.e. a fully fledged Arabic system). The following section will further describe these two system types.

4.3.1 Arabised systems

In the last two decade the Arab world has witnessed a number of attempts to introduce automated systems to their environment (i.e. libraries, information centres, offices etc.). There are a few Arabised systems used in Arab libraries such as DOBIS/LIBIS, MINISIS, and STAIRS. The problems and stages of Arabisation of these systems have been discussed by Ashoor, (1989); Booth *et al,* (1986); Al Dosary and Ekrish, (1991); and Al-Anzi and Collier (1994). These systems may handle both bibliographic and full text databases. For example, the Saleh Kamel Centre in Al Azhar University uses the MINISIS system for storing and retrieving the full text of 16 books of Prophet Moahmmed traditions. However, most libraries and information centres use these systems for bibliographic records. The Saudi Software Company has launched the Arabisation of the Bibliographic Retrieval System (BRS) for storage and retrieving Arabic full text documents in Arabic (Saudi Soft, 1995).

This approach is relatively easy to implement at the price of abandoning some Arabic language characteristics. The command system needs to be Arabised and consequently an I/O interface must be built (Al Kharashi, 1991). However, this method has not been successfully used in Arabic language processing (Abu Salem, 1992). This may refer to the nature of the Arabic language (i.e. the writing system from right to left; the structure of the word and sentences in Arabic).

Briefly, there are several problems that have been noted by those who are involved in the Arabisation of non-Arabic systems (Booth *et al*, 1986; Ashoor, 1989; Morfeq, 1990). These problems can be summarised as follows:

- Arabised systems may lack the ability to represent all the symbols and characters of the Arabic script,
- the Arabic language is written from right to left while Latin languages are written from left to right,
- the Arabised systems do not consider any linguistic approaches to the Arabic language,
- the Arabised systems do not allow the use of Arabic diacritical marks,
- the Arabic language is completely different from Latin languages in terms of both its written system and the form of the Arabic letters which are joined up as in مكتبة (library).

Some of the above problems faced the current researcher when he was working at King Fahd National Library (KFNL). Generally speaking, some of the above problems can be avoided if more attention is paid during the period of Arabisation. In brief, the problems or difficulties related to Arabisation can be categorised into two types:

- those related to the people who are involved in Arabisation, and
- those related to the system itself.

For example, the DOBIS\LIBIS system was Arabised (unfortunately) by three different academic organisations at the same place (Saudi Arabia). Differences can be noted between these versions of the system. This is due to the fact that one organisation may used more features which are available in the system than the others. Regarding the problems related to the system itself, some systems offer more features that may help in handling Arabic than others.

However, the Arabised systems still suffer from a shortage of features that make it capable of handling Arabic in an efficient way. This is due to the fact that these systems do not consider a linguistic analysis to the language. This point is most important and should be kept in mind during the design of the information or text retrieval system in Arabic. The present thesis will pay particular attention to dealing with linguistic analysis for text retrieval, especially in the area of morphological analysis.

Before leaving this section, an important point about Arabisation should be addressed. In the researcher's personal opinion, though Arabisation is a significant task for improving information technology in the Arab world, it is not the whole solution. Arabisation may have been very significant in the past when there were few skilled peoples in the Arab world, and technology was not in an advanced state but nowadays the situation is different; there are thousands of Arab scientists in this field (i.e. computing, and information technology etc.) who can carry out this task (i.e. to design complete systems in Arabic), and to gain the benefits of global technology, adapting it to fit in with the Arabic language.

4.3.2 Fully Fledged Arabic systems

Given the problems and difficulties of successfully Arabising Latin or other non-Arabic systems, a number of companies and organisations have attempted to design and develop fully Arabic information systems. The underlying motivation is based on the belief that the systems which were designed to handle non-Arabic characters would not succeed in handling Arabic characters, as mentioned previously. For example, in 1993, the ASSET company launched IRSAD (Computer Guide, 1993), a system for the full text retrieval of documents (Information Retrieval System for Arabic Documents). This offers most of the standard techniques typically present in English information retrieval systems, such as logical operators, proximity searching and simple field-based record structures. The IRSAD system could be run on Windows, Novell Netware or Unix SCO.

Another system is AFTDB (Arabic Full Text Data Base) designed by Al Alamiyah of the Electronic Company (Sakhr Software, 1997). This is the result of five years' work on computational linguistics in Arabic. The AFTDB is a powerful system for handling Arabic data. It has been applied on the Quranic and Al Hadith texts. It offers a number of search features such as morphological searching, lexical text searching, proximity searching and Boolean searching. In addition, there have been a number of Masters and Doctoral projects which have attempted to design Arabic information retrieval systems. Examples of such systems include the Bayan text database management system (Morfeq, 1990); the Micro-AIRS system (Al Kharashi, 1991), and the thesis by Al Naim (Al Naim,1989) under the title

"Text analysis and automatic indexing for Arabic based automated information retrieval system".

4.4 Information retrieval evaluation

According to the literature in English, a number of comparative studies have been carried out to investigate the comparison between titles or document abstracts and full texts. For example, "a series of tests for automatic indexing was conducted with three document collections. One collection out of three, a set of eighty-two short papers presented at the 1963 Annual Meeting of the American Documentation Institute, was used to test a full text of documents with thirty-five search requests. The comparison between the document abstracts (an average of fifty-nine words in length) and full texts (an average of 1,380 words in length) was done in two fields of processing: a word stem dictionary and synonym dictionary (thesaurus). The results show that full text processing is superior to abstract processing in both word stem and synonym dictionaries. But Salton and Lesk point out that the increase in effectiveness is not great enough to reach the unequivocal conclusion that full text processing is always superior to abstract processing" (Tenopir and Soon Ro, 1990).

Tenopir's study (Tenopir and Soon Ro, 1990) evaluated full text searching compared with searching on different fields in the Harvard Business Review Online database on the BRS search system using thirty-six queries from the history file of two libraries. Their conclusion was that full text searching had higher recall and retrieved a greater total number of documents, but it had lower precision than abstract or controlled vocabulary searching.

As for Arabic information retrieval, a few experimental studies have been carried out to investigate which method is most suitable for Arabic. A study by Al Kharashi (Al Kharashi, 1991) describes how he built a microcomputer-based Arabic information retrieval system to compare the use of full words, stems, or roots as index terms. He conducted a series of experiments of 355 documents with titles only. His study reveals a superiority of root and stem retrieval methods over the word retrieval method. The root performs as well or better than the stem at low recall levels, and definitely better at high recall levels.

A similar study was done by Abu Salem (1992) which repeated Al Kharashi s experiment comparing the use of words, stems, and roots as index terms taking 120 titles in computing with their abstracts. Abu Salem found that the stem retrieval method performs significantly better than the word retrieval method, while the root retrieval method performs significantly better than the word retrieval method. At lower recall levels (up to 0.6), the root retrieval method does not perform significantly better than the stem retrieval method. However, at higher recall levels, the root retrieval method performs significantly better than the stem retrieval method. These two previous studies compared the three methods for the retrieval of titles or abstracts only. Furthermore, the samples of the two studies were small; Al Kharshi took 355 titles, and Abu Salem only 120 titles with their abstracts.

Hmeidi (1995) designed an automatic indexing system for information retrieval in Arabic. He also tested three methods of search: namely, full word, stem, and root as index terms. Hmeidi's results, as he mentions, confirm the results of Al Kharashi and Abu Salem with smaller corpora; that root indexing is more effective than word indexing. The sample which was used in the Hmeidi study was 242 abstracts in computing and information systems. An important point should be noted here about the studies of Al Kharashi and Abu Salem; this is that both studies used manual indexing rather than using stemming or root finding algorithms (automatic indexing) in order to create the inverted file (index terms file). Of course, the output of both methods of indexing (manual and automatic) may be different. In other words, the performance of each method is different. The present study is plans to use automatic indexing that is, to create the index terms, rather than using manual indexing.

Morfeq (1991) designed a retrieval system called Bayan, which can handle both Arabic and English text. The Bayan system is based on morphological analysis. Unfortunately, there is not enough data about the theories used for morphological analysis. In addition, there is no evaluation for the system mentioned in the study. Morfeq says that the search structure of the Bayan is not yet complete.

Al Naim (1989) designed a text retrieval system which was given the name Al Buhkary. This name is based on the whole text of Sahih Al Buhkary (about 7000 traditional texts of the prophet Mohammed) and this was used as a database. The Al Bukhary system is based on morphological and lexical analysis, the former being the heart of the system. The morphological analysis is based on the theory that divides Arabic words into two types. morphological balance words (80%), and unbalanced words (20%), each of which can be divided into subgroups. Al Naim (the designer) claims that the system is fast, accurate, and easy to use. However, no proof was given in the study to support the above statements Furthermore, Al Naim's study did not reveal what search methods the system can handle, and the performance of these methods.

Kaseem (1988) in his paper discussed the particularities of Arabic nouns and adjectives and their effect upon information retrieval. Kaseem dealt with the nouns and adjectives in a general sense without taking into account the needs of Natural Language Processing (NLP). Using nouns and adjectives for information retrieval is based on the fact that these forms are the most common types of word used in Arabic (Al Khuli, 1982). Furthermore, nouns are most often, if not always, used for indexing and subject headings (Al Atram, 1989). The above view was supported by the fact that most of the information retrieval evaluation which was done by (Al Kharashi 1990; Abu Salem, 1992; Al Atram, 1989; Al Sawydan, 1993) used nouns and adjectives as search queries.

4.5 Natural Language Processing for Arabic (NLP)

In the last two decades it has been noted that particular attention has been paid to linguistic analysis for Arabic. Though NLP in Arabic is still in its infancy, it can be noted that a number of conferences and meetings have been held in the Arab world to discuss problems and present potential solutions. There are also a number of companies which are involved in this area. For example, the Al Alamiyah computational linguistics group led by Dr N. Ali embarked, in 1985, on a long-term research project to develop an understanding of Arabic text with the following major objectives:

- To disambiguate Arabic written text both morpho-syntactically and lexically,
- To provide extensive support to Arabic lexico-graphical efforts,
- To support advanced research in stylistics, content analysis (mainly, automatic abstracting and indexing, and text generation) (Ali, 1992).

Current NLP in Arabic, as Hegazi and El Sharkawi (1986) pointed out, divides Arabic linguistic analysis into four stages and these can be listed as follows:

1- Morphological analysis, which associates features of various sorts with words, taking no account of context. This thesis deals with morphological analysis. This is due to the fact that this type of analysis is the most important for information retrieval. Furthermore, the remaining levels need in depth analysis to find out how the language behaves.

2- Syntactic analysis, which associates syntactic structures with phrases, taking into consideration the respective positions of words.

3- Semantic analysis, which aims at removing the ambiguities remaining after syntactic analysis, by reference to the semantic relations which bind concepts together.

4- Pragmatic analysis, which places the phrase in the context of the general realm of knowledge in order to remove the ambiguities which cannot be eliminated by semantic analysis (Hilal, 1985).

El Sadany and Hashish (1989) refer to the late application of computational linguistics in the Arab world in the following observations:

- until recently there has been little interaction between computer scientists and Arabic linguists. At first, most of the systems dealing with the Arabic language were developed by engineers and computer scientists. Thus, the systems developed performed only small demonstrations that ran only to a few examples collected by the system designers without facing the real problems of the Arabic language itself.

- because of the nature of the Arabic language it needs special techniques and algorithms for solving its morphological problems. Unfortunately, most of the previous work in the field of Arabic morphology has applied techniques and algorithms used with western languages. This point was also noted by Anwar (1989), who stated that Arabic computational linguistics draw on different theories that were designed outside Arabic studies; these are mostly one or more versions of Transfer Grammar (TG), Generalised Phrase Structural Grammar (GPSG) or Functional Grammar (FG).

To a certain extent, I do agree with the above statement but, at the same time, I would like to say a few words here about the linguistic theories used in English or European languages. In my view, a good deal of benefit can be gained from these theories on condition that the structure of Arabic should kept in mind when these theories or techniques are used. Furthermore, some of theories are universal, and this means that they can be used in any human language with some modifications. Examples of such theories are: Augmented Transition Networks (ATNs), Definite-Clause Grammars (DCGs), and Finite-State Transition Networks (FSTNs). For example, the ANTs have been implemented, with some modification, on a Multi-Mode Morphological processor (MMMP) which was designed by Al Alalmiyah, and used with a Quranic and Hadith database (Ninth of the books of Al Hadith) in order to extract stems from given terms by removing prefixes and suffixes attached to the given terms. In fact, the English language and other European languages were served by Western linguists, especially regarding those issues related to computational linguistic theories, while on the other hand, the Arabic language suffers from the absence of computational linguistic theories. As a result of that it might be helpful to adapt Western linguistic theories to serve Arabic.

Current work in Arabic Natural Language Processing (NLP) suffers from a lack of co-ordination between various groups in the Arab world, which leads to repetition of work and has created non-standard research environments (Al Jabri and Mellish, 1994). An excellent piece of work which paid equal attention to both sides of computational linguistics (i.e. language and computing) can be found in Ali (1988).

4.6 Morphological analysers for Arabic

In the last two decades, a number of morphological analysers have been developed for various purposes. They differ in many ways. Some of the ways in which they differ are outlined in the following sections. Generally speaking, the existing Arabic morphological analysers can be broadly divided into two types or categories: the linguistic approach and the non-linguistic approach (i.e. mathematical approach). The linguistic approach can be further subdivided into a number of classes. Al-Uthman (1989) attempted to classify the existing morphological analysers in the following way:

1- *the khaleeliah rule* (prefix and suffix removal). The idea of this method is to remove the prefixes and the suffixes that may be attached to the root. A list of prefixes and suffixes are needed. This method may fail to treat those words which allow infixes in them. An example of this method is the Hilal system (Hilal, 1985).

2- *hierarchy based on morphological and phonological rules.* This approach uses the morphological derivation rules that have 3, 4, 5, 6 or 7 characters, and the syllabic patterns of each word. This is also a pattern matching method, but it is represented in a different way. The method is based on morphological and phonetical rules. The morphological rules are represented by all the Arabic morphological patterns which number about 400 patterns while the phonetical rules are represented by syllabic patterns. Hegazi and El Shorkawi's system belongs to this category.

3- *pattern matching and flags.* This method uses a list of Arabic patterns which number about 400. A given word is checked against these patterns in order to reduce it to its root. Thalouth and Al Dannan's analyser is an example of this method.

4- *combinatorial approach.* This is also a type of pattern matching presented in a mathematical way. This approach is based on identifying the positions of the root's letters in a given word by examining patterns. For example, given the word خطير *khater* (danger) which contain four letters, all patterns with a length of four letters are examined. Suppose that these patterns are (فاعل, فَعيل). From the pattern فَعيل it is known that the root's letters are in positions one, two and four. Now the root of the word itself is known to be (خطر) and the pattern is (فَعيل) which gives the word(خطير). This approach requires little knowledge of Arabic roots. An example of this approach is Al Fedaghi and Al Anzi's algorithm.

In addition to the above classification, the following classes could also be added:

5- *two-level finite-state analysis.* This method is based on Koskenniem's Two-level Morphology. This approach postulates two distant but interrelated levels of representation for words. The alphabet used in Two-level Morphology is in fact a set of character pairs, called concrete pairs', each of which consists of a lexical character and a surface character which is one of its possible surface realisations.

6- *morpho-synthetic analysis*. This approach is based on language structure. It takes into account all matters that may affect the word structure such as affixation, morphology, phonology, syntax etc. The main aim of this approach is to reduce a given word to its root and identify all the forms that may occur in the dictionary related to it. MMMP is an example of this approach.

Generally speaking, there are a number of common steps which are shared by most morphological analyser algorithms and these can be summarised as follows:

- ❑ The largest prefix and suffix must be identified and removed to produce a word stem,
- ❑ The pattern of the word stem is then checked against all allowable patterns,
- ❑ If there is a match between the patterns, then the root of the word is generated and checked against a stored root list,
- ❑ If the root of the word is found in the root list then the process ends by returning the root, pattern, suffix and prefix,
- ❑ If there is no match between the roots, different patterns will be tried,
- ❑ If all patterns produce no root, then different combinations of prefixes and/or suffixes will be suggested to produce a different word stem (Al Kharashi, 1991).

The following section will give a brief review of morphological analysers which have been reported or published in the literature to date.

4.7 Examples of Arabic morphological analysers

4.7.1 Hilal's system

Hilal based his analysis on the three classes: a class of the largest prefixes, a class of the largest suffix and a third class using the number of letters remaining in the word). Hilal classifies the Arabic word into tools and ordinary words. Ordinary words follow grammatical rules. However, the tools do not follow such rules. The method of extracting triliteral roots from an ordinary word follows the general steps given below:

1- the longest possible prefix and post fix is eliminated by comparing the leading and the trailing characters with known prefixes and suffixes,

2- depending on the length of the remaining part, different rules are examined:

- ❑ eliminate extraneous letters,
- ❑ amend an extra letter to form a triliteral root,
- ❑ change a letter to its original value.

Many look-up tables are used to accomplish the task, including tables for patterns, roots, prefixes, suffixes and non-standardised roots (Hilal, 1985; Al Fedaghi and Al Anzi, 1989).

4.7.2 Hegazi and El Shorkawi

Hegazi and El Shorkawi designed a computer-aided morphological hierarchy (CAMH) system for vowelised Arabic text (Hegazi and El Shorkawi, 1985). The system is based on two rules: morphological and phonetic rules, and is able to derive the root of a word, its morphological balance and its morphological category. The system uses the morphological derivation rulers that have 3-7 characters and the syllabic patterns of each word. The system deals with about 400 patterns, which largely covers all the patterns in the Arabic language. A syllabic balance is done for every morphological pattern.

4.7.3 Thalouth and Al Dannan

This analyser (Thalouth and Al Dannan, 1989) used pattern matching and flags to implement differences in morphological rules. Moreover, prefixes and suffixes were defined in such a way that their interference with patterns was minimised. The algorithm in their system involved stripping the most frequent suffixes and prefixes, and then matching the word against a set of frequent patterns so as to obtain the triliteral origin of the word. The algorithm divided the words into three categories, each with its own treatment:

1- Structural word analysis which involves the following steps:

a- removing the largest prefix from the given word,

b- the remaining part will be checked against the structural word lexicon or pattern, c- if a structural word is found, the number of flags will be checked to specify the following:

- class of the accepted prefixes,
- whether the structural word needs to be analysed as a content word,
- whether the structural word is a compound, then the location of the components' boundaries,
- the grammatical value of the structural word.

2- Content word analysis

After stripping the prefix during the structural word analysis, the next analysis is content word analysis which involves:

- removing the long suffix from the given word,
- the remaining part is checked against patterns,
- for every pattern match, a number of flags are checked in order to eliminate unacceptable patterns,
- determine the possible roots tied to the word,
- check the existence of each possible root in the roots' lexicon,
- for each possible root which is matched in the lexicon, check if the original pattern is among the accepted patterns generated from this root.

3-Foreign word analysis

Foreign words mean words that have been widely adopted in Arabic. This analysis is invoked after the content word analysis. The following steps are carried out:

1- the foreign word lexicon will be checked,

2- the flag in this field will be checked to see if the prefix and suffix are accepted ones,

3- if step 2 fails, steps 1 and 2 will be repeated, first with the addition of the suffix, then with increasing parts of the prefix without the suffix.

Al Dannan and Thalouth's algorithms can be used as a reference for categorisation, translation and for other educational and cultural purposes. The algorithms have been implemented and tested. The developers of these algorithms claim that the results are quite encouraging.

4.7.4 Al Fedaghi and Al Anzi

Al Fedaghi and Al Anzi developed an algorithm (Al Fedaghi and Al Anzi, 1989) to examine every combination of three letters in the given word and produces its root-pattern representation. Two files are used in this algorithm as input; the file of triliteral roots and the file of patterns. The algorithm examines those patterns that have the same length as the input word. The main concept in this algorithm is to examine patterns in order to identify the positions of the root's letters in a given token. These letters are then tested to decide whether they form an Arabic root or not. The algorithm has been tested in four modes as follows:

- Mode 1: the input word contains its full triletral root,
- Mode 2: the third letter of the root of the input word is lost,
- Mode 3: one of the letters of the root of the input word is missing,
- Mode 4: two of the letters of the root of the input word are missing.

The test data involved several Arabic texts in order to represent different backgrounds. They compared each mode in terms of its speed and the capability to reduce a word to its root. They found that Mode Four took more than one hour to reduce the word to its root, while the other modes took between one and five minutes. With regard to reducing a word to its root, Mode One achieved 63%, Mode Two achieved 70% and Mode Three achieved 79%.

4.7.5 Bessley

Bessley *et al* (1989) and Bessley (1998) developed a morphological analysis algorithm based on Koskenniem's Two-level Morphology to identify all of the Arabic words with a valid morphological analysis. Two-level Morphology is so named because it postulates two distant but interrelated levels of representation for words. The alphabet used in Two-level Morphology is in fact a set of character pairs, called 'concrete pairs', each of which consists of a lexical character and a surface character which is one of its possible surface realisations. Concrete pairs are represented with the lexical character above the surface character, or, especially for computer work, with the sequence lexical_character : surface_character. A reserved null or empty character, here zero, can be used to show that a lexical character is not realised at all on the surface (zero realisation) or that a surface character can appear without any direct lexical counterpart (ex-nihilo realisation) as in the following example:

The Arabic word *albintu* (girl or daughter) can be presented as follows:

a- Vertical representation:

lexical character: #{al_bint+u#

surface character (and zero): 0@0l0b0nt000

b- Horizontal representation:

# : 0 lexical hash mark realised as nothing

{ : @ lexical eliding hamza (=a) realised as bar alif

a : 0 lexical short (a) realised as nothing

l : l lexical l realised as surface l

_ : 0 lexical _ realised as nothing

b : b lexical b realised as surface b

and similarly for i:0, n:n, t:t, +:0, u:0.

4.7.6 Multi-Mode Morphological Processor (MMMP)

In this section, more details will be given for the Multi-Mode Morphological Processor (MMMP). This is due to the fact that the MMMP is used for text retrieval in Arabic while the others are not (To the best of the researcher's knowledge to date). Furthermore, the MMMP is the analyser most related to this work in terms of the method of analysis and the use of the analyser for text retrieval.

The MMMP is a morphological analyser-synthesizer of Arabic words which are written from right to left. The analyser is based on analysis-by-synthesis. It is capable of analysing any Arabic word token regardless of its level of dicritisation into its derivational, inflectional, affixes and case ending primitives (Ali, 1992). The MMMP algorithms are designed so that they could be and adapted to different applications such as text analysis, text retrieval, automatic translation etc. The MMMP was developed in 1986 by the Al Alamiah Electronic Company in Cairo, and in the subsequent year (1987) it was incorporated into a Full Text Data Base for Arabic (AFTDB); and, as Ali explains (Ali, 1988), it was implemented to compress and search the full text of the Quran on a morphological basis. The AFTDB was used in 1989 to develop an Al Hadith database Since then the Al Alamiah Group has used the AFTDB as a database engine to produce several other Arabic full text databases such as *The Encyclopedia of the Nine Traditional* books, *Fiqh Al mua`amalat* etc.

At this point it is important to mention that the MMMP algorithms are based heavily on linguistic analysis. Generally speaking, linguistic knowledge, especially morphological analysis and word decomposition processors, are at the heart of the processor. The theoretical analysis of morphology was carried out by a linguistic group working with the Al Alamiah Company in order to achieve a more accurate analysis.

The MMMP can be used as both a morphological analyser and a generator To generate a word from out of its primitive constituents, the processor works in the opposite direction to the morphological analyser. As was mentioned above, the analyser works on the principle of "analysis-by-synthesis". This method, as Ali (1988) asserted, is to be used and would be helpful when numerous linguistic ambiguities occur in a language (such as Arabic).

The developer of the MMMP designed the analyser to be able to hypothesise a number of treatments for a given word, by analysing the word and then resynthesizing it. If a match is found between the input word and the analysed word, then the treatment is regarded as correct. If not, the analyser should repeat the same process with a new treatment. Briefly, the main function of MMMP is to provide all valid diacritical forms for the input word. For each valid diacritisation, it provides the morphological decomposition of the word (affixes, stem). The stem is further analysed in terms of its root and morphological forms. The MMMP is divided into four parts as follows:

- ❑ Morpho-Syntactical Processor
- ❑ Derivational Processor
- ❑ Parsing Processor
- ❑ Diacritisational Processor

4.8 Morphological analysis and information retrieval in Arabic

As was mentioned earlier, the Arabic language is an inflectional and derivational language which means that, given a single root, one can derive hundreds of words. For this reason, Plessis (1990) points out that morphological analysis is an absolute necessity for any computer processing of the Arabic language. A frequency list of a document cannot be compiled before an analysis of the words is made. If a program was to take words as they occur in the text, a word such as ولد *wld* (boy) would be counted under ولد *wld* (a boy), والولد

walwld (and the boy), الولـد *alwald* (the boy), and in numerous other places. This would not give a proper indication of the number of times that the word ولد *wld* (boy) was used in that text. Al Naim (1989) emphasises the importance of using morphological analysis for the Arabic language for the following reasons:

1- it is estimated that some 80% of Arabic words are derived from generative roots; either triliteral, quadrilateral, or pertaliteral verbs,

2- the number of generative roots are few enough to be easily processed by typical computers of today,

3- the root-derived words are usually generated by either modelling the root, attaching prefixes and suffixes, or by applying both. Therefore, by knowing the morphological structure of the word, one can easily trace it back to its root and vice versa. Moreover, such prefixes and suffixes are limited to only 15 and 25, respectively,

4- The most frequently used morphological forms do not exceed 400 which makes it even easier to process such forms.

Prefixes and suffixes cause many problems for text retrieval in Arabic. To overcome these problems, Al Kharashi (1991) suggests that the system should be designed to strip out all suffixes and prefixes from every extracted index term before adding them to the inverted list, while Al Bakhit (1993) believes that if the left, right and infix truncations are available in the system these problems will be resolved by using these truncations.

However, simple truncation, whilst important for English language retrieval, is less applicable to languages such as Arabic which rely heavily on affixation. Furthermore, truncation may be used wrongly by the user. There are two errors that may occur when truncation is used by the user. These are:

1- over-truncation. This type of error occurs when too short a stem remains after truncation; this may result in completely unrelated terms being conflated to the stem. For example, suppose that a user is searching for the term ***catalogue*** and at the search time he/she truncates the term as cat* (* this sign used to show where the truncation is located). Of course, in addition to the word(s) ***catalogue/s***, there will be a number of completely unrelated words that will be retrieved such as ***categorise, cathedral, cattle,*** and so on.

2- under-truncation. This happens when too short suffix is removed and this may result in related words This type of truncation will miss some relevant words For example, if a user is looking for texts or documents related to computers, and he/she truncated the term as ***computer*** only, words such as ***computing*** and ***computational*** will not be retrieved.

Having said that truncation may not bė an appropriate technique for Arabic, there is an alternative technique which has been used to support Arabic information retrieval systems instead of using truncation. This technique is known as a morphological analyser which is capable of reducing the variety of a given term to its root, automatically, in order to retrieve all the variations of that term.

Generally speaking, word stemming or morphological analysis performs two useful functions in information retrieval system, as was mentioned by Lennon *et al* (1981). Firstly, it reduces the total number of distinct terms present, with a consequent reduction in dictionary size and updating problems. Secondly, similar words generally have similar meanings and thus retrieval effectiveness may be increased if conflation is carried out on both document and query terms. The following points are a summary of the benefits of using word stemming for information retrieval:

1- stemming makes it possible for a user to retrieve morphologically related terms which may also have a semantic relationship,
2- one way of broadening the search in information retrieval is to use word stemming,
3- by using word stemming the recall factor will be increased.

Though there are these advantages in using word stemming, sometimes there is a risk in using this tool, such as the retrieval of unrelated words.

4.9 Summary

The main goal of an information retrieval system is to search an information repository and retrieve records that are potentially relevant to a query. As Tague-Sutcliffe (1996a) stated: achieving this goal has always been a core activity of information professionals and has been a concern of information science since the field arose in the 1950_s Information retrieval performance is influenced by a number of factors. Language is among the factors which has an effect on information retrieval. Another major problem with IR is information growth or overload, which has become one of the most significant problems facing information retrieval users whether using online databases or the World Wide Web

In Arabic information retrieval system, there are two approaches to the design retrieval system. The first approach is to Arabise an existing system from another language (usually English) into a format which is capable of handling Arabic text.. The second approach is to build an Arabic system to handle Arabic text. As for Arabic information retrieval, a few experimental studies have been carried out to investigate which method (word, stem or root) is most suitable for Arabic. In the last two decades it has been noted that particular attention has been paid to linguistic analysis for Arabic. Though NLP in Arabic is still in its infancy, it can be noted that a number of conferences and meetings have been held in the Arab world to discuss problems and present potential solutions. Current work in Arabic Natural Language *Processing (NLP) suffers from a lack of co-*ordination between various groups in the Arab world, which leads to repetition of work and has created non-standard research environments.

Morphological analysers for Arabic have been developed for various purposes. They differ in many ways. Some of the ways in which they differ are outlined in the following sections. Generally speaking, the existing Arabic morphological analysers can be broadly divided into two types or categories: the linguistic approach and the non-linguistic approach (i.e. mathematical approach). The chapter gave a brief review of morphological analysers which have been reported or published in the literature to date. Prefixes and suffixes cause *many problems for text retrieval in Arabic. Some author (Al Kharashi, 1991)* suggests that the system should be designed to strip out all suffixes and prefixes from every extracted index term before adding them to the inverted list, while Al Bakhit (1993) believes that if the left, right and infix truncations are available in the system these problems will be resolved by using these truncations. However, simple truncation, whilst important for English language retrieval, is less applicable to languages such as Arabic which rely heavily on affixation. Having said that truncation may not be an appropriate technique for Arabic, the alternative technique is known as a morphological analyser which is capable of reducing the variety of a given term to its root, automatically, in order to retrieve all the variations of that term.

Chapter Five: Experimental design and methodology

5.1 Introduction

This chapter describes the research methodology adopted in the study, starting with the experimental design. The experimental design is divided into two parts: the experimental environment and experimental procedures. In the first part the test cases are discussed, in addition to the study variables which may have an effect on information retrieval performance. The latter part deals with the procedures followed whilst running the experiments. The chapter ends with the retrieval performance measures that are used in this study.

5.2 Experimental design

Generally speaking, the problem in information retrieval is to assess the effects of one or more factors, or independent variables, on a performance measure by means of a sample of experimental units in which each unit is assigned to a combination of factor levels and measured as to performance. Experimental design is concerned with techniques for assigning the experimental units to the factor levels. In order to eliminate bias, it is essential that units be assigned randomly, and in order to assess interactions between variables, it is essential that more than one unit be assigned to combinations of factors (Tague-Sutcliffe, 1996b). The current study aims to evaluate the retrieval performance of four search methods: namely, word, stem, root, and morpho-semantic methods. Each method is used to search different queries under various groups of documents.

As noted by Tenopir, Nahl-Jakobovits, and Howard (1990) using experiments on information retrieval has the advantage of controlling extraneous variables, producing quantifiable results, and such experiments can be undertaken by a single researcher. The main disadvantage is that it does not directly involve end users. It provides no information on how an end user interacts with a full text database and system and how that system helps his/her request for relevant documents. The general representation of the experimental design of the study is shown in Figure 5.1 below. Further details of each element of the experimental design will be discussed in the following sections.

Figure. 5.1 the experimental design representation

Independent	variables	Dependent Variables	
Treatment Condition	Treatment Group	Outcome	
		Recall	Precision
Word Stem Root Morpho-semantic	Group 1	O_{1-4}	O_{1-4}
	Group 2	O_{5-8}	O_{5-8}
	Group 3	O_{9-12}	O_{9-12}
	Group 4	O_{13-16}	O_{13-16}
	Group 5	O_{17-20}	O_{17-20}

O: Observation

As can be seen from Figure 5.1, the entire study sample (treatment groups) is divided into five groups. Each group is used to search four methods (treatment conditions). This treatment aims to show how a single search method behaves under each unit or group.

5.3 Experimental environment

It is known that there are a number of variables and factors which may affect the retrieval performance of an information retrieval system; examples include the test collection, user needs, subject coverage etc. Therefore, attempts were made in the present experiments to control as many of these variables as possible. Further details of controlling the variables of the current study are discussed later in this chapter. Before discussing these variables, a description of test data collection is given.

5.3.1 Test collections (Samples of the study)

Information Retrieval (IR) experiments often use test collections, which consist of a document database and set of queries for the database for which relevance judgments are available. The number of documents in test collections has tended to be small, typically a few hundred to a few thousand documents (Frakes and Baeza-Yates, 1992). Much traditional work on IR systems' evaluation has taken place using one or more of the traditional test collections. The existence of portable test collections (with queries and relevance judgments) has been a substantial factor in the development of research in the field (Robertson and Hancock-Beaulieu, 1992).

Test data collection for English texts are readily accessible. However, for Arabic, unfortunately, there is no complete test collection which can be used by the Arab information retrieval investigators, although there are a number of attempts to fulfil this task. King Abdulaziz City for Technology and Science (KACTS) test collections are an example of uncompleted test collections. The next sections are descriptions of the test collections of this study.

5.3.1.1 General view of the Sample

A judicious choice of the study sample has been made. Before choosing the sample, a number of characteristics were considered. These characteristics can be summarised as follows:

- the sample should be written in Arabic language,
- it should be a text or abstract or title,
- the sample should be a written passage (not spoken) without dertical marks,
- it should be in Arabic standard style,
- it should contain new terms which have been introduced into Arabic, such as computer, Internet and so on,
- it should be in fluent Arabic, not slang,
- the sample should not contain poems,
- the sample should be collected from a variety of sources such as newspapers, magazines, periodicals, books etc,
- the samples should cover a variety of subjects,

As Salton and McGill (1987) mentioned:

> For many practical purposes, it is sufficient to use document excerpts for analysis, such as titles and abstracts. The available experimental evidence indicates that the use of abstracts in addition to titles brings substantial advantages in retrieval effectiveness. However, the additional utilization of the full texts of the documents appears to produce very little improvement over titles and abstracts alone in most subject areas.

5.3.1.2 Size and source of the sample

The study sample (test collection) was gathered from the following sources:

- ❑ KACTS test collections (CD-ROM),
- ❑ KFNL Abstracts Journal (floppy disk),
- ❑ Morshd Indexes (CD-ROM).

1- KACTS test collections

KACTS test collections contain 30 MB of Arabic texts without queries or relevance judgments. The KACTS test collections are saved on one CD-ROM. It was prepared by Dr. Ibrahim Al Kharashi. He is a specialist in Arabic information retrieval evaluation; he works with KACTS in Riyadh, Saudi Arabia. The KACTS test collections have been collected from a variety of sources as follows:

- ❑ daily newspapers,
- ❑ weekly magazines,
- ❑ academic journals,
- ❑ monthly journals,
- ❑ conference proceedings,
- ❑ internet sites,
- ❑ books,
- ❑ encyclopedias,
- ❑ electronic publications.

Dr. Al Kharshi and his team have been collecting these samples for the purpose of information retrieval evaluation. Unfortunately, no queries or relevance judgments were developed for the KACTS collections. The language of the texts and abstracts is the Arabic standard written language used in books, newspapers and other publications. Generally speaking, the language of KACTS can broadly be categorised into two types:

- General language, which is used by the majority of Arab writers,
- Scientific language, which may be used in the field of science.

However, the collectors at KACTS aimed to cover a range of the subjects which might differ from one another in terms of linguistic styles. The subject coverage of KACTS can be divided broadly into the following subjects:

- Politics,
- Environment,
- Information Science,
- Education,
- Economics
- General topics (Islamic Studies, Law, History, Literature etc),
- Science (Computer, Mathematics, Chemistry, Biology).

2- KFNL Abstract Journal

The KFNL Abstract Journal is one of King Fahd National Library's (KFNL) publications. It is published four times a year. It contains about 300 abstracts. The majority of the abstracts are in the field of Library and Information Science. The journal is available in electronic format.

3- Morshd Indexes

The Morshd indexes contain two types of indexes:

- A book index
- A journal index

Both indexes contain 350,000 entries. The study uses only the journal index, which covers about 600 Arabic journals. The number of articles which were indexed is 11,600. The main coverage of this index is Islamic Studies and Arabic Literature.

5.3.1.3 Choosing the study sample

As was mentioned above, the population of the study is too large, and it is beyond the scope of this work to use the whole collection to run the experiments. For the purpose of this study, a small sample suffices. Therefore a method was developed to choose between 500-600 records from the three sources (population). Whilst choosing the sample, more than one method was used. For the selected test sample to be representative, the following methods of selection were adapted:

- for some samples the choice was based on random selection;
- for some samples the first fifty records were selected;
- for some samples all the records were selected;
- for some samples a semi-random selection was applied.

It is meant by semi-random sample is that from the sample of *Morshd Indexes*, the journal index was selected, then about twenty index terms were used to search the index. Then all the records relating to these index terms were used as a sample for this study.

In choosing a sample from Science and Technology, the selection was made an original sample of 1015 tables of contents. The sample was numbered from 1-1015. The first 300 records were chosen as a primary sample of the study. From the 300 records, a third of these sufficed. The following steps show how the random numbers were used to draw a random sample from the primary sample (300 records). Since the number of samples needed was 100 records, primary sample of 300 was divided by the needed sample 100. The number 3 results. This number became the first member of the sample. To get the second member of the sample, number three was added to the first member of the sample, to get number 6 as the second member of the sample. To get the third member, again number three was added to the second member to get number 9 as the third member of the sample. Number three is added to each member of the primary sample until the number 300 is reached. This is the last member of the sample.

5.3.1.4 subject coverage

After finishing the selection procedure, some 590 records were collected. This is the main sample of the study is used to run the experiments. The 590 records were divided into 5 groups according to subject coverage. Figure 5.2 gives details about the subject coverage and the number of records for each subject.

Figure 5.2 sample of the study (Test collections)

Collection	Type of documents	Subject	Records	Queries
Group 1	Article titles	Economics	237	8
Group 2	Journal Abstracts	Library and Information	85	6
Group 3	Full text Articles	General Articles	79	3
Group 4	Book content tables	Science and Environment	148	12
Group 5	Proceeding abstracts	Education	41	3
Total			**590**	**32**

5.3.2 Variables

Retrieval performance is influenced by a number of factors. For example, Schamber (1994) listed about 80 factors that have been found, or have been suggested in the literature, to affect relevance judgments only. In experimental, observational, or survey studies, the researcher tries to understand the relationships between the variables involved. Particularly in experiments, researchers look for changes in the *dependent variables* that occur as a consequence of change in the *independent variables*. Variables affecting retrieval performance are numerous. Some variables are related to user needs or satisfaction, others are related to subject coverage of the domain, searcher experience or searching strategy etc. Generally speaking, variables affecting retrieval performance can be categorised broadly into the following:

- Independent Variables,
- Dependent Variables,
- Extraneous Variables.

The variables of the study are categorised as above. The three categories of the variables affecting information can be further divided into six components, as suggested by Tague-Sutcliffe (1996b). The six components of an information retrieval performance are shown in Figure 5.3. Under each component several variables can be listed.

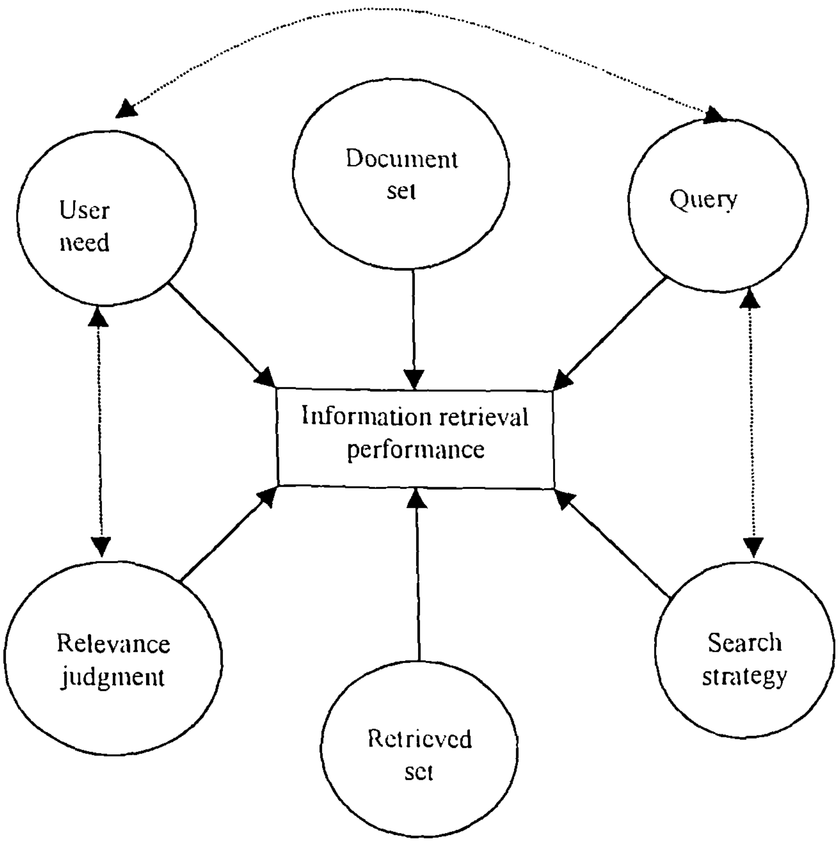


Figure 5.3 variables affecting information retrieval

5.3.1 Independent Variables

Researchers are interested in identifying the effects of selected independent variables on elements of the search process and its outcome; they choose to investigate certain independent variables because they believe that these variables have some effect on the search process and/or its outcome (Fidel and Soergel, 1983). The independent variables of the study are the search methods. This study looks at how a given method affects the search outcome or the retrieval performance. The following methods are tested as independent variables:

- word method,
- stem method,
- root method, and
- morpho-semantic method.

The first three methods are used in today's Arabic information retrieval systems, while the fourth method is introduced by this study. The retrieval performance of the fourth method will be compared to the other three methods, and the methods will be compared to each other.

5.4.2 Dependent Variables

The dependent variables in information retrieval evaluation usually relate to the search process and/or the search outcome; for example, speed of searching and number of databases searched are variables characterizing the search process. Much would be gained by identifying clearly these variables, and by knowing what variables might affect them The dependent variable of this study is the search outcome. This variable can be further broken into: the outcome in terms of the recall measure and the outcome in terms of precision. Further details of the two measures are discussed in Section 5.6.

5.4.3 Extraneous (uncontrolled) variables

Extraneous variables are those variables that have an effect on information retrieval performance (such as indexing policy, cost), and these will not be tested by the current study. However, although these variables are not tested here, some of them may have indirect effect on the retrieval performance, and will need to be controlled as much as possible. The extraneous variables which will be controlled in this study are described in the following sections.

5.4.3.1 Word forms

Because of morphology (word formation), a word may appear in a set of documents in variant forms (noun, verb, adjective, adverb etc). Furthermore, each of these forms may also appear in different forms (in the case of a verb, it can appear in the past, present, or future; it can be passive or active and so on). The verb and adverb forms will be controlled (i.e. will not be tested here), and the only forms to be tested in the study are nouns and adjectives as independent variables.

5.4.3.2 Synonyms

Synonyms mean different words with the same, or nearly the same, meaning (e.g. *purse* and *handbag*) (Fromkin and Rodman, 1998). This study will not use all the synonyms of the key words of queries. Because of that, this type of processing is related to semantic analysis rather than morphological analysis and so searching for synonyms of keywords is beyond the scope of this study.

5.4.3.3 Subject coverage

The subject coverage of the database has an effect on the retrieval performance. This type of variable is discussed above, when the sample of study is described.

5.4.3.4 User needs

Though a rich literature of user needs or satisfaction exists in the information retrieval context, yet the debate about the issues surrounding the topic is still on-going. Tague-Sutcliffe (1996b) summarised the situation in the following statement:

> To evaluate how effective the system, some writers believe that the original user must be involved in the relevance judgments. Others believe that at least some aspects of a system can be evaluated without relevance judgments from the users. Relevance judgments, in this view, represent judgments of whether or not the document is about the query, and so can be made by any knowledgeable person. The other view of relevance judgements is that they represent the value of the document for a particular user at a particular point in time and so can be made by the user only at that time.

The current study will not involve real users and, because of this, the main aim of the experiments are to evaluate the effectiveness of the performance of the four methods of search, not users and their information seeking behaviour. Therefore, the study developed queries and search statements instead of using real queries or real search problems. The following section deals with this issue.

5.4.3.5 Queries (search statement)

To avoid confusion about the query or request, it is helpful to cite the definition of these terms, which was given by Robertson (1981). According to Robertson, "A query or request is a statement by a/the requester describing his/her information need, but recently it has come to mean simply the act of requesting". Queries may be divided into two types:

- real-life, which represent real information needs of users, or

- artificial, which may be derived from titles or other parts of the document (e.g. titles, abstracts, or index terms).

The latter type was used in this study to avoid the complexity of user information needs and their information seeking behaviour. Sometimes there are strong reasons for constructing artificial queries rather than acquiring real ones because many investigators find real users difficult to control, difficult to involve in the search process and therefore difficult to evaluate in a predetermined fashion. Users do not willingly participate in tests, they will drop out before completing all the requirements, particularly the evaluation of the output, and will not obey instructions for maintaining the integrity of the experiment (Tague-Sutcliffe, 1992).

Having opted for artificial queries it is important to give a brief description of their characteristics:

- All the queries are single or multi-word statements. "Statement" in this study refers to those words or terms which will be used to search the database,
- All the queries were derived from a sample of the study,
- All queries were identified by experts in the field to which the sample belongs.

As mentioned above, the study sample was divided into five groups. From each group a number of queries was derived. The total number of queries was 32.

5.4.3.6 Search strategy

The search strategy has a great impact on information retrieval performance. In brief, the search strategy can be defined as "the plan of a search for information, involving specification of needs, choice of search terms, degree of specificity in searching, how to extend the search to broader, narrower and related classes" (Buchanan, 1976). In fact, there are a number of variables that may affect the search strategy, such as the experience of the requester using the system, or the requester about the subject, or using the Boolean operators AND, OR, and NOT for broadening or narrowing the search output.

As was mentioned above, the main aim of the study is to evaluate retrieval performance of the search methods (i.e. word, stem, root and morpho-semantic methods) Therefore, the following search strategies were adapted:

- The only Boolean operator used during the search was the AND operator as a conjunction of two terms or more;
- Truncation is used normally to remove a prefix or suffix of a given word in order to retrieve all word variations. Since this type of strategy is related to morphological variations of the word, the morphological analysis is capable of searching for the morphological variations of the given word on behalf of the user;
- Word synonyms will not be used during the search;
- This researcher carried out all the searches. As mentioned above, one of the advantages of using experiments on information retrieval is that a single researcher undertakes the searching tasks.

5.4.3.7 Relevance judgment

It is not the aim of the current study to give a lengthy review of relevance. The aim is to give a brief summary of this issue, and define relevance in the context of the study Further details about work on relevance can be found in Saracevic (1975), and Schamber (1994).

Relevance may be used to refer to various relationships (i.e. to information needs, topic, or information request) or it may be used in different ways. It has been examined by information scientists throughout the history of the field (Schamber, 1994). Despite a rich literature on relevance, there is no commonly accepted definition. Although much experimental research in IR is based on the idea of relevance, various interpretations of the concept have been made (Park, 1993). Relevance is also a confusing and much debated concept. In the context of this study, relevance is used to refer to a relationship between a document and a request statement (i.e. matches between subject terms in queries and subject terms in documents.)

Swanson, in his article on relevance (Swanson, 1986), divided relevance into two types: subjective and objective relevance. Subjective relevance expresses a relationship between retrieved documents and the user; the user is the final arbiter of what is relevant.

Objective relevance expresses the relationship between a document and a request statement. In other words, objective relevance expresses the logical link (or topical link) between the query and the document. Relevance, in this sense, is a connection between a written request and a document and belongs to the world of objective knowledge (Park, 1993). Furthermore, relevance in this sense also tells nothing about the degree of success that is achieved in meeting the information needs of users, as Lancaster and Warner (1993) mentioned. A given document, or set of documents, is relevant to a query, or queries, no matter who the user is or what the user expects from the search (DeSalvo, 1992). The relevance in this study will be called objective relevance or topical relevance.

Having defined the meaning of relevance as used in this work, there is a question which needs to be answered. Who is qualified to make the relevance decision? It might be the information specialist associated with the system, the requester, or an independent subject specialist. Clearly, the person making the decision must know enough about the subject matter to be able to ascertain that certain documents are a "legitimate response" to the request and others are not (Lancaster and Warner, 1993). In fact there are a number of factors that may affect the judgment. Schamber in her excellent article, mentioned some of these factors (Schamber, 1994):

- the most important characteristic of judges is the degree of subject knowledge or related professional education,
- the more judges know about a given subject area, the more their judgments tend to agree,
- the more judges know about a subject area, the fewer documents of a retrieval set they tend to judge relevant,
- judges who are most involved with the problem tend to agree in their judgments,
- judges tend to agree on relevance rankings of documents in a set regardless of their relevance ratings of individual documents,
- Judges tend to agree more on judgments of "non-relevant" than on judgments of "relevant".

5.4.3.8 Judges

In this study, subject specialists assigned the relevance judgments. They are research students in different subjects, studying in U.K. universities. As was mentioned previously in this study, relevance means the topical link between the query and the texts in the eyes of the judges. However "if a judge decides that a certain record is sufficiently close in subject matter to a particular request the system was correct in retrieving it" (Lancaster and Warner, 1993). The researcher students were asked to give their opinion about the query and the records. The judges were given instructions to consider the queries and rate the records as relevant or not relevant based on their understanding of the queries and the records obtained.

5.5 Experimental procedure

5.5.1 Running the experiment

In order to run the experiments of the study, a prototype system was developed. The description of the system can be found in Chapter 7. Running the experiments involved the following:

1. All the queries were divided into five groups according to the sample groups,
2. Each group of queries were subjected to the following methods:
    - □ word method,
    - □ stem method,
    - □ root method, and
    - □ morpho-semantic method.

The search procedures continued until all the group's queries were searched.

5.5.2 Saving results

Following the completion of each search, the output of each method was saved in a file. Saving the results in files is useful for further data analysis.

5.6 Retrieval performance parameters

A number of retrieval performance parameters are used by information retrieval researchers as a basis for comparing the features of different systems or different search methods. In fact, a number of parameters are used to investigate the effectiveness of the retrieval outcomes. It is known in the information retrieval systems' evaluation that most records are assumed to be classified as either relevant or not relevant, for a given query using binary *RSV* (retrieval status value). Then, records are either retrieved or not in response to the query. A two-by-two matrix or contingency table (2x2 table) is formed to represent the *RSV,* as in Figure 5.4, where rows represent the number of records retrieved (or not retrieved) and columns represent the number which is relevant (or not). The letters a-d represent the number of records with the indicated attributes; e.g. a is the number that was both relevant and retrieved. Based on this, a number of parameters can be defined (Boyce *et al* 1994).

	Relevant	**Not Relevant**	**Total**
Retrieved	a	b	a + b
Not Retrieved	c	*d*	c + *d*
Total	a + c	b + d	a + b + c + d

Figure 5.4 A 2 X 2 table of retrieval outcomes based on binary relevance measures.

In most information retrieval systems, when a search is conducted, the system divides up the document collection into two parts. The documents that match the search strategy used to interrogate the system are retrieved (a+b), and all the documents that fail to match the strategy are not retrieved (c+d). This dichotomous partitioning of the document collection may be regarded as a form of system relevance prediction. As based on the 2 x 2 table, a number of parameters can be defined. This study uses the following parameters:

5.6.1 Recall and precision measures

Recall and precision are the most important and well-known measures in the field of information retrieval evaluation. Recall ratio measures how well the system retrieves all the relevant records, and it is computed as:

$$\textbf{\textit{Recall}} = a/(a + c)$$

Precision ratio measures how well the system retrieves only the relevant records, and this is computed as:

$$\textbf{\textit{precision}} = a/(a + b)$$

5.6.2 Noise or false drop and omission factor

The noise or false drop used as the ratio of irrelevant records retrieved to total retrieved. It is computed as:

$$\textbf{\textit{Noise or false drop}} = b/(a + b)$$

The second parameter is the *omission factor*, which is defined as the portion of relevant records not retrieved. It is computed as:

$$\textbf{\textit{Omission factor}} = c/(a + c)$$

5.6.3 Failure analysis

Failure analysis for recall and precision is used in this study in order to understand why recall and precision failures occurred and what the reasons are for these failures.

5.7 Summary

This chapter outlines the experimental design, which has been carried out by the study. It is known that there are a number of variables and factors which may affect the retrieval performance of an information retrieval system; examples include the test collection, user needs, subject coverage etc. Generally speaking, the variables can be categorised broadly into: independent variables, dependent variables, and extraneous variables. The independent variables of the study are the search methods (i.e. word, stem , root, and morpho-semantic method). The dependent variable of the study is the search outcome, while the extraneous variables are those variables that have an effect on information retrieval performance (such as indexing policy, cost) but not tested by the study.

A detailed description of the study sample was represented in the chapter. The retrieval performance parameters (mainly recall and precision) used in this study were also covered The next chapter discusses the framework of the study which will be implemented later on

Chapter Six: The morphological system in Arabic

6.1 Introduction

The morphological system in Arabic plays a major role in word formation. This in turn will affect the retrieval performance in Arabic. Therefore, a deep discussion of Arabic morphology is essential. This chapter deals with ιwo types of morphology in Arabic, known as derivation and inflection, each of which has systematic rules for word formation. The chapter discusses issues most related to this study.

6.2 Morphology

Morphology has a great impact on word formation. This section describes the nature of morphology. In linguistic study, morphology is concerned with the forms of words themselves. Most linguists agree that morphology is the study of the meaningful parts of words. The elements of meaning in words can be perceived. For example, in the English word *walked*, two elements of meaning are present: *walk* plus *-ed* (indication of past tense). *Walk* and past tense are generally called morphemes (a unit of grammar smaller than the word). In the word *books* two morphemes *book* and plural are present. A word such as *unhelpful* has three morphemes: negative + *help* + adjective. But terms such as negative, plural, and adjective are abstract; they are not real forms. The real forms that represent them are called morphs (the smallest sequence of phonological units into which words are divided in an analysis of morphemes). Thus, the above words will be divided as follows (Crystal, 1991b):

Words	*Morphs*	*Morphemes*
walked	*walk-ed*	*walk* + past
books	*book-s*	*book* + plural
unhelpful	*un-help-ful*	negative + *help* + adjective

The morph which can stand as a word on its own, such as *book*, is called a **free morph**, while morphs which cannot stand as words on their own, such as *ed*, *s*, or *ful* are called **bound morphs**.

6.3 The Arabic Morphological system

At the beginning of this section it seems that it is important to discuss three terms related to morphology in Arabic. These terms are: morphology, inflection and derivation. There is no complete agreement amongst Arab linguists about the use of these terms. Arab linguists have used these terms for centuries. Al Asa'ad (1993) gives a historical background about the use of these terms by Arab linguists. Some linguists used the term morphology as a synonym of derivation, while some others use the term inflection as a synonym of derivation. However, most, if not all, the Arab linguists agree that morphology is the study of word formation. In the context of this study, the three terms are used and defined as mentioned in chapter 1.

Arabic morphology

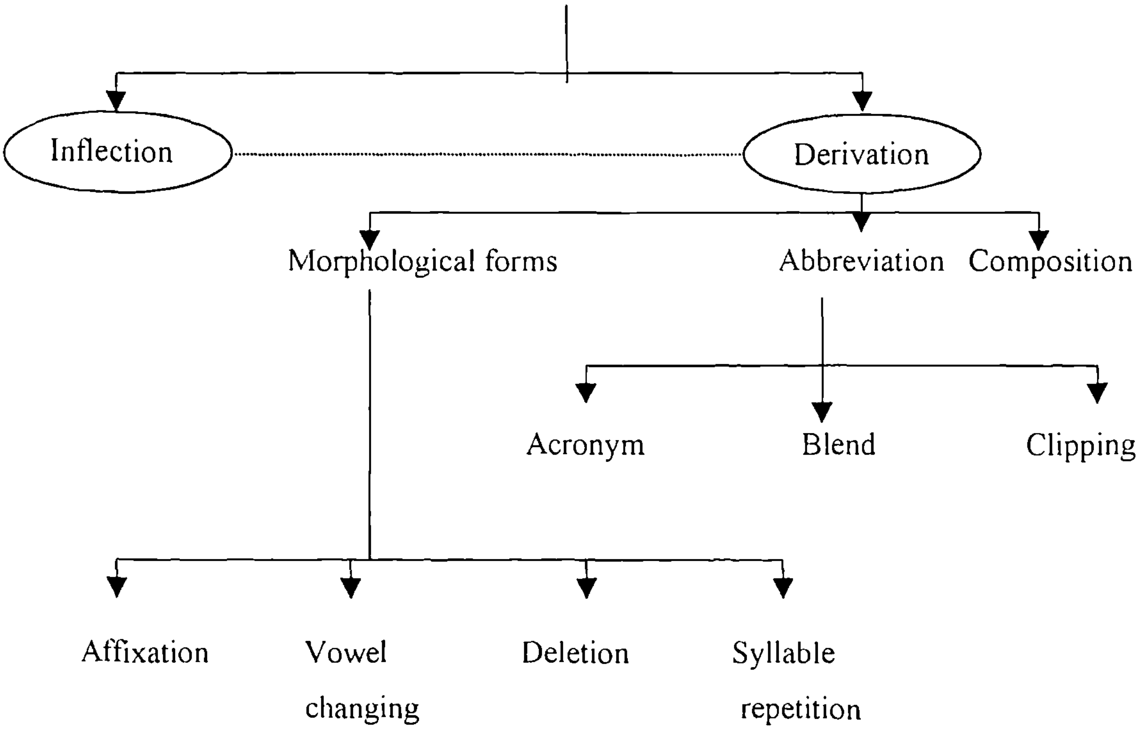


Figure 6.1 Arabic morphology in Arabic

· As shown in Figure 6.1, there are two types of morphology in Arabic: derivation and inflection. Derivation consists of three types morphological forms, abbreviation and composition. Fortunately, abbreviation and composition are very rare in Arabic It is clear from the above figure that, there is a strong relationship between derivation and inflection One should note that the inflectional morphology usually comes after the derivational process in Arabic. The structure of Arabic morphology can be represented as shown below in Figure 6.2.

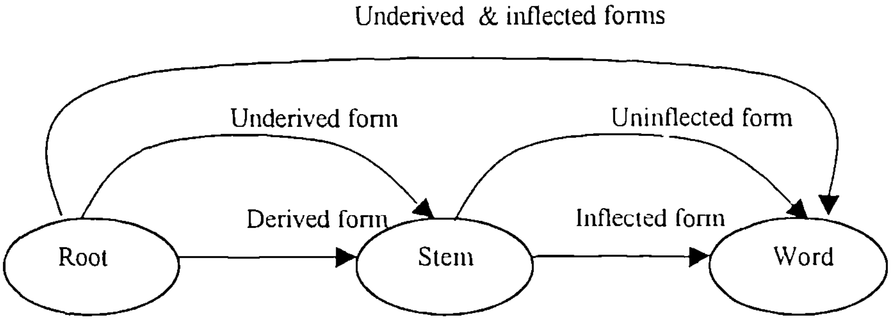


Figure 6.2 A broad structure of Arabic morphology

As can be seen from Figure 6.2, the derived form is generated through derivational morphology, while the inflected form is generated by inflected morphology. However. an Arabic word could be one of the following:

- ☐ underived and uninflected word; or
- ☐ derived word; or
- ☐ inflected word.

6.4 Derivational morphology

In fact it is not easy to give a complete definition of this type of morphology. According to Crystal (1991b), derivation is a " Major type of word formation where a certain kind of affix is used to form new words. A contrast is intended with the process of inflection, which uses another kind of affix in order to form variants of the same word". As noted by Fromkin and Rodman (1998), students often ask for a definition of derivational morphemes. The two authors try to give a simple answer: "the derivational morphemes are bound morphemes that are not inflectional. Inflectional morphemes signal grammatical relations and are required by the syntactic sentence formation rules. Derivational morphemes, when affixed to roots and stems, change the grammatical word class and/or the basic meaning of the word". To simplify the definition of the derivational morphology, consider the word المدرسون *almodaresun* (teachers). This word is composed of four morphemes. The root is ***drs***, from which we derived the word ***modares*** (agent noun), and then the added definite article ***al*** as a prefix. If the inflectional suffix ***un is added,*** the plural in its nominative case would result. The hierarchical structure of this word is shown in Figure 6.3 below.

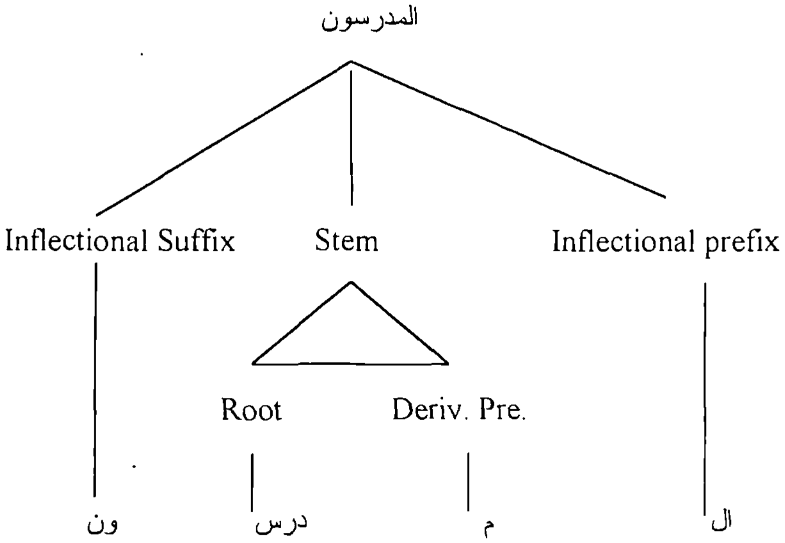


Figure 6.3 Structure of the Arabic word almodaresun (the teachers)

Derivation may change the category of the word, add semantic information, and/ or change the phonology of the base. The derivational morphology in Arabic composes two things: root and morphological forms.

6.4.1 Derivational methods

As shown in Figure 6.1, derivation in Arabic can be divided into three types: morphological forms, abbreviation, and composition. In the following sections further details of derivation-based morphological forms is represented, while the last two types of derivation are represented below:

1 Abbreviation

Derivation based on abbreviation can be achieved by:

- Syllable cut-off (for example, demonstration can be abbreviated as demo. or department as dept.)
- Omitting vowels (such as building: Bldg; boulevard as Blvd)
- Using the first letters of the words (such as United Kingdom: UK).

Methods 2 and 3 cause difficulty in Arabic, because they will break the morphological form of the word. Fortunately, they are not popular in Arabic. However, some authors do not consider the omission of vowels and use of the first letters of words to be within the morphology domain (Ali, 1998).

2 Compounding

Compounding can be defined as two or more forms which combine to build a new form. Word formation is not only based on a single word, but may be formed from two words as well as in, for example, the English words "mailbox" or "blackboard". An example in Arabic is قروأوسطي which is a combination of the two words القرون الوسطى (Middle Ages).

6.4.2 Derivation-based Morphological forms

This type of derivation in Arabic is the most important and recognisable type. The majority of Arabic words are based on this type of derivation. The processing of derivation-based morphological forms in Arabic is very complicated, and is the most difficult area of Arabic morphology. Arab morphologists have developed a powerful system of derivation. The system consists of several elements such as the root system, morphological form measures, and the derivational path. Further details of these elements are discussed in the later sections. In brief, derivation based on morphological forms can be subdivided into four methods as follows:

- derivation-based affixation (augmented letters);
- derivation-based vowel changing;
- derivation-based deletion;
- derivation-based syllable repetition;

The first method is the most common method used in Arabic derivation. However, all the above methods may occur in one word (Ali, 1988).

6.4.3 Arabic root system

A root is the base form of a word which cannot be further analysed without the loss of the word's identity; or alternatively, that part of the word left when all the affixes are removed (Crystal, 1991a). An Arabic root is an ordered sequence of three or four valid characters from the alphabet. However, the majority of Arabic words are derived from the bare root by adding vowels or any letter(s) of the set (سألتمونيها). See the next section.

The root has a general, basic meaning, which forms the basis of many related meanings. These related meanings are represented by the root consonants put in different forms *wazn* (measure or morphological forms) which differ in terms of the short vowels and other letters which are added (Owens, 1988).

As there are 28 letters in the Arabic alphabet, there is, in theory, the following number of possible permutations:

1- Bilateral: 784 $(28)^2$ roots; the actual number of used roots is only about 22.

2- Triliteral: 21,952 $(28)^3$ roots; the actual number of used roots is about 7,000.

3- Quadrilateral: 614,656 $(28)^4$ roots; the actual number of used roots is about 2700.

4- Pertaliteral: 17,210,368 $(28)^5$ roots; the actual number used is only about 300.

Although there are over 17 million possible roots in Arabic, the number of roots actually used is much less. There are about 10,000 different used roots (Wall, 1989). Through derivational morphology, one can generate millions of words through the composition of the 10,000 roots and the morphological forms. As Fromkin and Rodman (1998) noted, this method of morphology reflects the wonderful creativity of the language. Examples of derived words from a single root are shown in Table 6.1 below.

Table 6.1 Some derivatives from the root KTB (Smart, 1986)

Root	Arabic	Transliteration	Translation	Notes
*K-T-B (ك ت ب)	كتاب	KiTaaB	Book	Simply *a* choice of internal vowelling with no prefixes or suffixes, but infix (ا).
	كتابة	KiTaaBah	Writing	Feminine suffix added to the previous word to change its meaning.
	كاتب	KaaTiB	a writer, clerk	Change of internal vowelling.
	مكتب	maKTaB	office, desk	The very common prefix *ma* (m) here, plus another change in vowelling.
	مكتبة	maKTaBa	library, bookshop	Same as above, but again the feminine ending is used to change the meaning.
	مكاتبة	mukaaTaBa	correspondence	Another common prefix *ma* (m) plus another change in vowelling

*The root letters are given in capitals for the sake of clarity.

6.4.4 Morphological forms

The morphological form in Arabic is a system that, given an arbitrary word, defines the way this word can be used in the language. The form *fal* (to do) is used as a paradigm, whence the first radical of the trilateral verb is called by Arab morphologists *fa*, as the letter *f* in English, the second *ain* as *a* and the third *lam* as *l*. This is well known in the field of morphology as the *fal* pattern.

Every morphological form is consists of a root and a pattern (composed form). In order to generate an Arabic word from the root, one should follow the morphological forms. This can be done by adding vowels or any letter(s) of the set: (*sin, hamzah, lam, ta, mim, waw, nun, ya, ha, alif*), known in Arabic as augmented letters "سألتمونيها", to some part of the root. For example, the root *DRS* consists of three consonants (radical). It is pronounced in English as *D, R,* and *S*. Vowels or letters of the set are added to the root in order to derive words. or stems. The additional augmented letters or vowels are controlled by the derivational rules of Arabic. Let us take the above example (*DRS*), and see how a member of the augmented letters is added to the three radical letters of the root. The root *DRS* (which means to teach) is shown in Figure 6.4 at level 1. From this root we can generate several words by passing them to the morphological form system (level 2). The output of the derivational processing is a full Arabic word, as shown on level 3 of Figure 6.4.

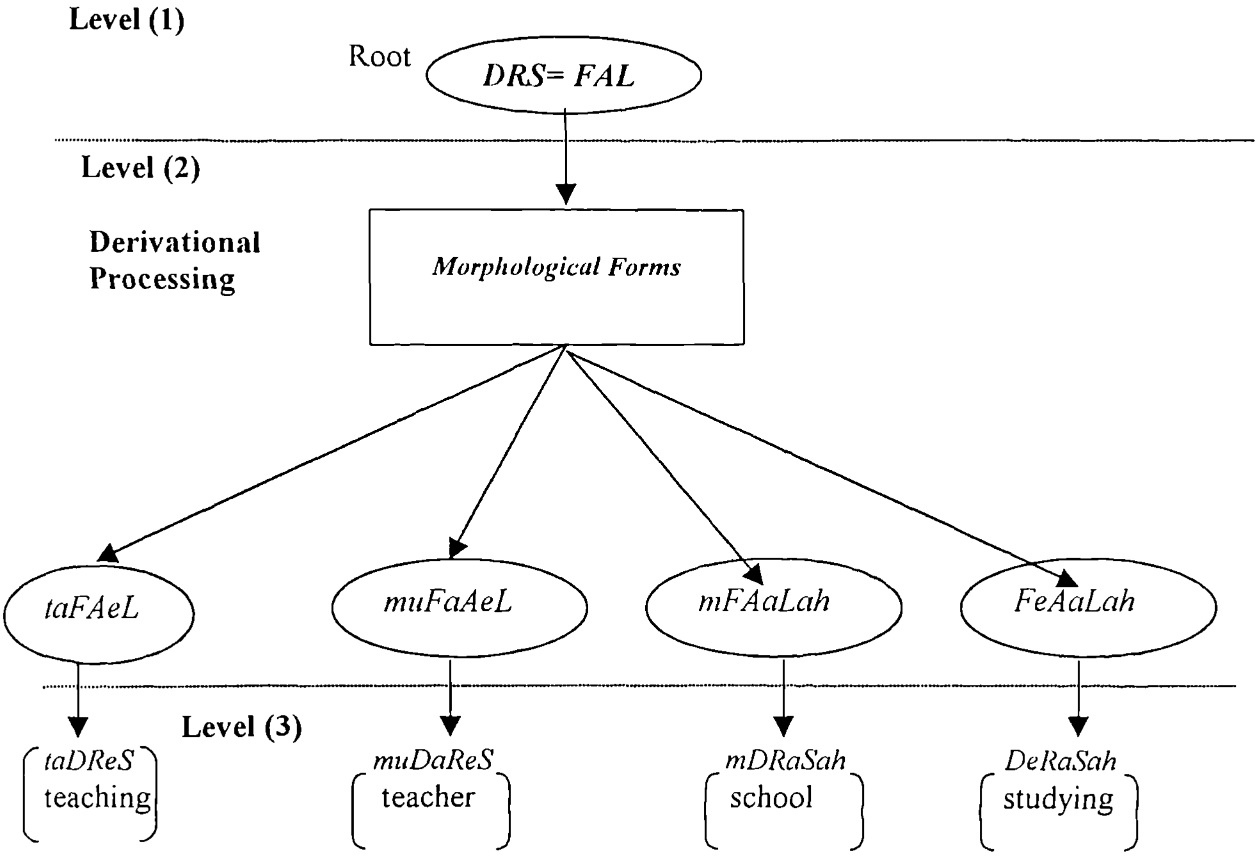


Figure 6.4 Word derivation based on morphological forms

In Arabic there are two types of morphological forms as follows:

1. Verb forms

The great majority of Arabic verbs are trilateral, that is to say, contain three radical letters; quadrilateral verbs are relatively rare. From the first or ground-form (form I as shown in Table 6.2) the trilateral and quadrilateral verbs are derived, in different ways, from several other forms, which express various modifications of the idea conveyed by the first. The derived forms of the trilateral verb are usually reckoned to be fifteen in number, but the last five forms are very rare occurrences (Wright, 1974).

Table 6.2 Morphological forms of derivatives (verbs)

FOR M.	*ARABIC*	*Translitration*	*SEMANTIC MEANING*
I	فعل	*FAL*	Simplest, starting point for further derivation.
II	فعل	*FAa'aL*	To do frequently or intensively, to consider somebody as...
III	فاعل	*FaAaL*	To direct, strive to, act in conjunction with ...
IV	أفعل	*aFAL*	To shape into ..., induce, cause to do ...
V	تفعل	*taFAa'aL*	To become ..., to do to oneself, to claim to be ...
VI	تفاعل	*taFaAaL*	To act mutually, to simulate.
VII	انفعل	*inFaAaL*	To allow action be done to oneself; reflexive.
VIII	افتعل	*eFtaAaL*	Reflexive of I; may be used instead of VI or VII
IX	افعل	*aFAuL*	To be or become a certain colour, or marked by a certain defect.
X	استفعل	*estaFAaL*	To ask somebody for something, to force oneself, to do unto oneself; reflexive of IV

Through these forms several forms can be generated via a morphological path, as will be discussed in the following sections.

2. Noun forms

There are two types of noun in Arabic: derived nouns, and underived nouns. The derived nouns, like verbs, are derived from the root. The derived nouns are semantically related to the root. There are nearly 400 morphological forms of derived and underived nouns (El Sadany and Hashish, 1989). The morphological forms of the derived nouns can be divided into broad categories (as shown in Table 6.3), each of which may be subdivided into a subclass.

Table 6.3 Morphological forms of the derivatives (nouns)

NUMBER	*DERIVATIVES*	*ENGLISH MEANING OF DERIVATIVES*
1	اسم الفاعل	Agent noun (Active participle)
2	اسم المفعول	Passive participle
3	الصفة المشبهة	Qualificative adjective
4	صيغ المبالغة	Forms of exaggeration (intensiveness)
5	اسم التفضيل	Comparative noun
6	اسم الزمان	Noun of time
7	اسم المكان	Noun of place
8	اسم الالة	Instrumental noun
9	المصدر	Verbal noun

For example, the instrumental noun (a derived noun used to refer to the means by which an action is performed) has seven forms. All the derivatives which were listed in Table 6.3 are derived from verb forms which are listed in Table 6.2. In order to generate a word from a root in Arabic, one must follow the derivational path. Further details of derivational paths in Arabic are discussed in the next section. The main function of the morphological forms is their use in generating new words or analysing them.

6.4.5 Derivational path

In the previous sections the root system and how the morphological forms work in Arabic have been discussed. It was also mentioned that if one wants to generate or analyse an Arabic word one should use the morphological forms. The morphological and generating analysis is based on using the derivational path. The derivational path consists of several rules such as rules of derivation, rules of morphological forms and phonological changing. Figure 6.5 shows an overview of the relationship between derivatives. Further relationships are given in Appendix 4.

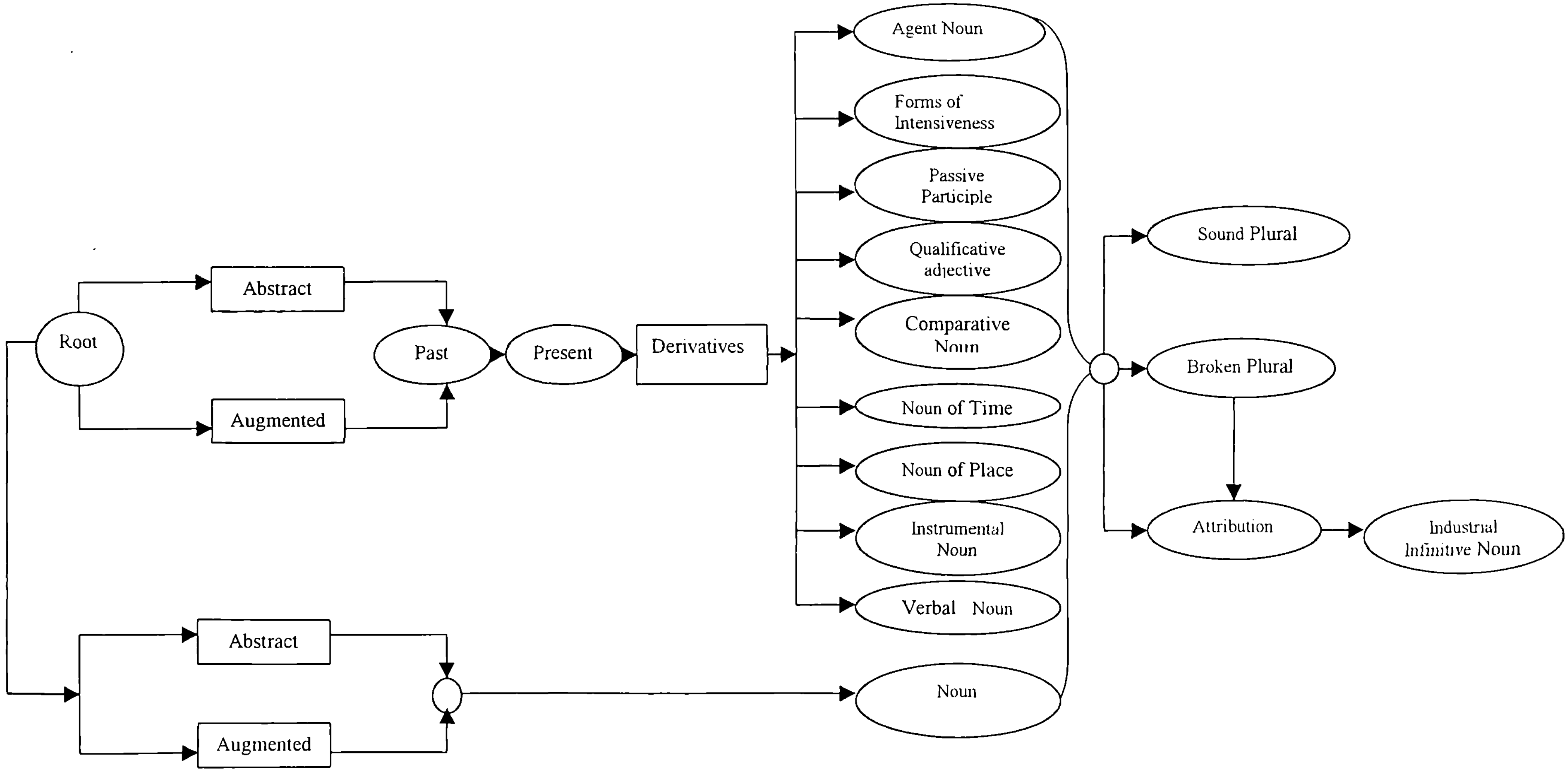


Figure 6.5 Derivational path in Arabic

However, before discussing the derivational path in Arabic, it is necessary to raise an important point about the source of derivation in Arabic. In Arabic literature related to derivation, there is no complete agreement about the source of derivation, whether a word itself (Ali, 1988), a verb or verbal noun (Daraz, 1986), or a root as suggested by the Arab lexicographers. Yaqub (1982) quoted from Tarzy stated that there is not just a single source for derivation. He added that, generally speaking, all derivatives might be reduced to verbs Owens (1988) gives a summary of this issue by saying at the end of his discussion on derivation that:

> I should perhaps not leave this subject without mentioning one of the more famous questions in Arabic linguistics, whether the noun is derived from a verb or a verb from a verbal noun. I think it fair to say that the argument is pretty much a stand off. Methodologically the issue is not of great importance, and so when, for example, *Ibn Jinni* (a well-known scholar of morphology) talks about derivation from a root he can simply refer to derivation from a specific root without specifying whether the derivation is from a verb or a verbal noun.

This study will use the root as a source of derivation as suggested by the Arab lexicographers. This choice is influenced by the following facts:

- most Arabic words, whether nouns or verbs, can be reduced to a single root,
- the root is the base form of a word which cannot be analysed further without loss of the word's identity,
- Arab morphologists use the form *FAL* (three radical characters) as a paradigm of the Arabic root. This pattern equals the majority of Arabic roots,
- roots in Arabic play a major role in derivation,
- using the root as a source of derivation can be applied automatically and more easily than other theories (i.e. verb or verbal noun),
- the majority of Arabic lexicons are based on root arrangement.

Having identified the source of derivation in Arabic, the time has now come to deal with the derivational path. First of all, one question should be answered. What do we mean by a derivational path? In summary, the derivational path in Arabic is a systematic method of representing all the morphological forms and the relationship between them. In other words, the derivational path tells us which form is derived from which form. For example, consider the derivation of a passive participle noun. In fact, there are two types of passive participle noun:

- □ abstract passive participle
- □ augmented passive participle

Each of these has a specific rule. In this example, it is necessary to look at the derivation of these forms. To derive them, there are a number of morphological rules. These are as follows:

RULE 1:

IF the verb is abstract, triliteral, and in the position of the passive past tense THEN the morphological form of the passive participle noun is (مَفْعُول) *maFAuaL* (node 4 as in Figure 6.6).

RULE 2:

IF the verb is augmented and in the position of the passive present tense THEN change the first character of the verb to (م) (Arabic character as *m* in English). One morphological form of this rule is (مُفَعَّل) *moFaA'aL* (node 6 as in Figure 6.6).

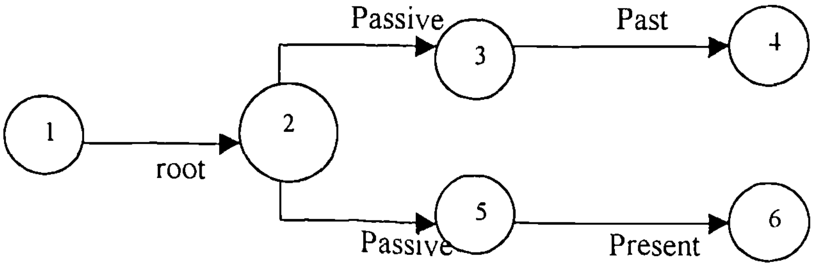


Figure 6.6 passive participle derivation

There are about 73 morphological forms of the passive participle noun in Arabic (Yaqub,1993). The above two rules can be applied to most of the morphological forms of the passive participle noun.

The function of the morphological rules is to control the derivational process through the derivational path. Figure 6.5 represents the most well-known morphological forms in Arabic and the relationship between them. Figure 6.5 also shows a general overview of the derivational path which can be broken down into a number of sub-paths, as shown in Figure 6.7, which represents the morphological path of the augmented verbal noun.

Suppose the root علـــم *ALM* (something relevant to the infinitive "to know") is taken. Starting from the node (ROOT) as in Figure 6.7 and moving forward, the past tense of the root, which is علـــم *AaLaMa* (taught) can be generated. Then, moving toward the present verb node, the verb يعل م *yoAaLiMou* (to teach) can be generated. Moving farther forward, the augmented verbal noun can be generated as in تعليم *taALeeM* (teaching).

The derivational path in Arabic is a very powerful system to be used in an automatic morphological analyser (or generator). Furthermore, by using the derivational path the inference mechanism can be used which enables the generation of new facts from existing ones by applying knowledge which has already been acquired to new situations. For example, the agent noun and verbal noun have the same sources of derivation. Therefore, the semantic link is rewritten between the two forms by using the inferring rule as follows:

FOR any X and Y,
X has a semantic link of Y, IF
(1) both X and Y have the same source of morphological form, and
(2) Z is the morphological form.

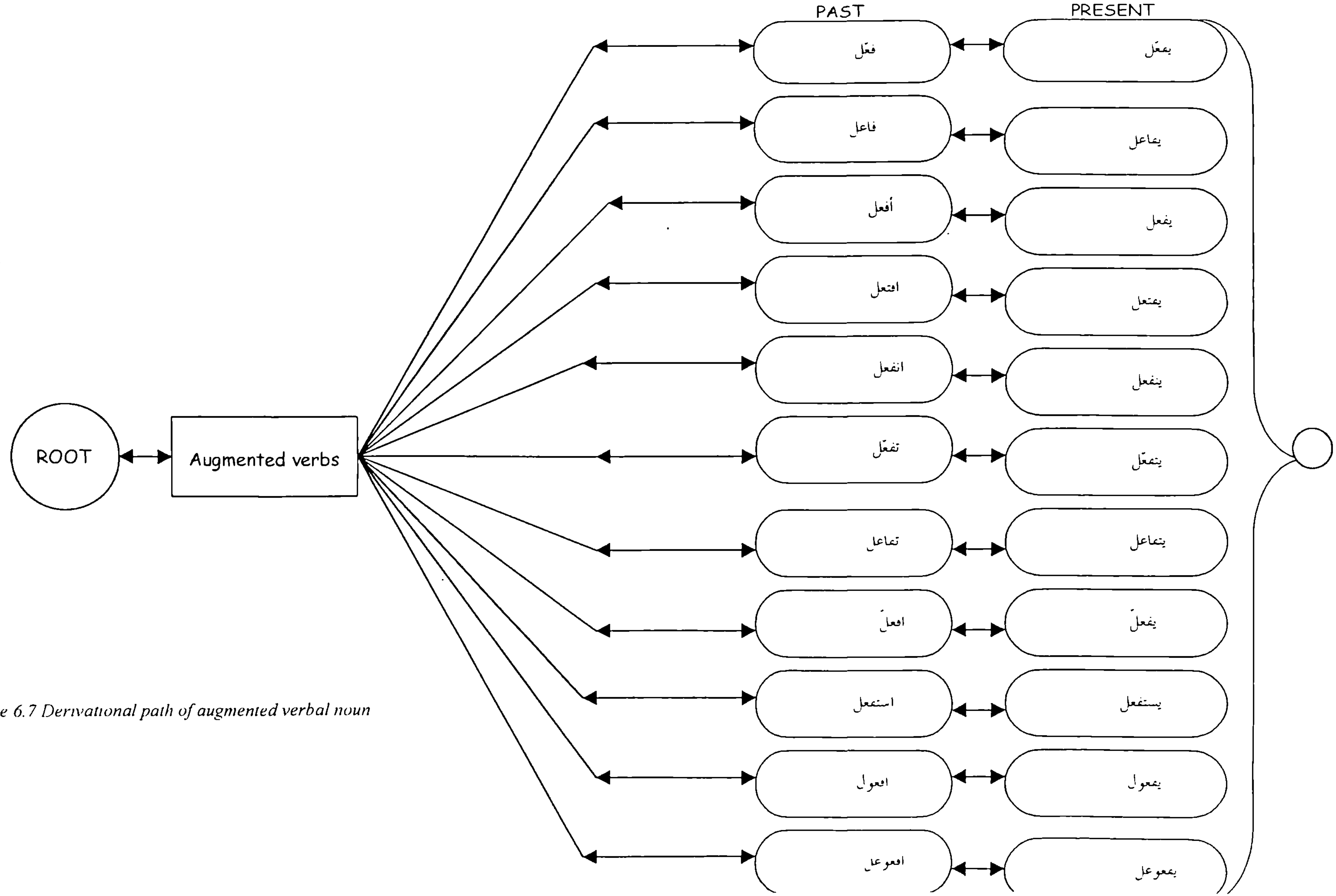


Figure 6.7 Derivational path of augmented verbal noun

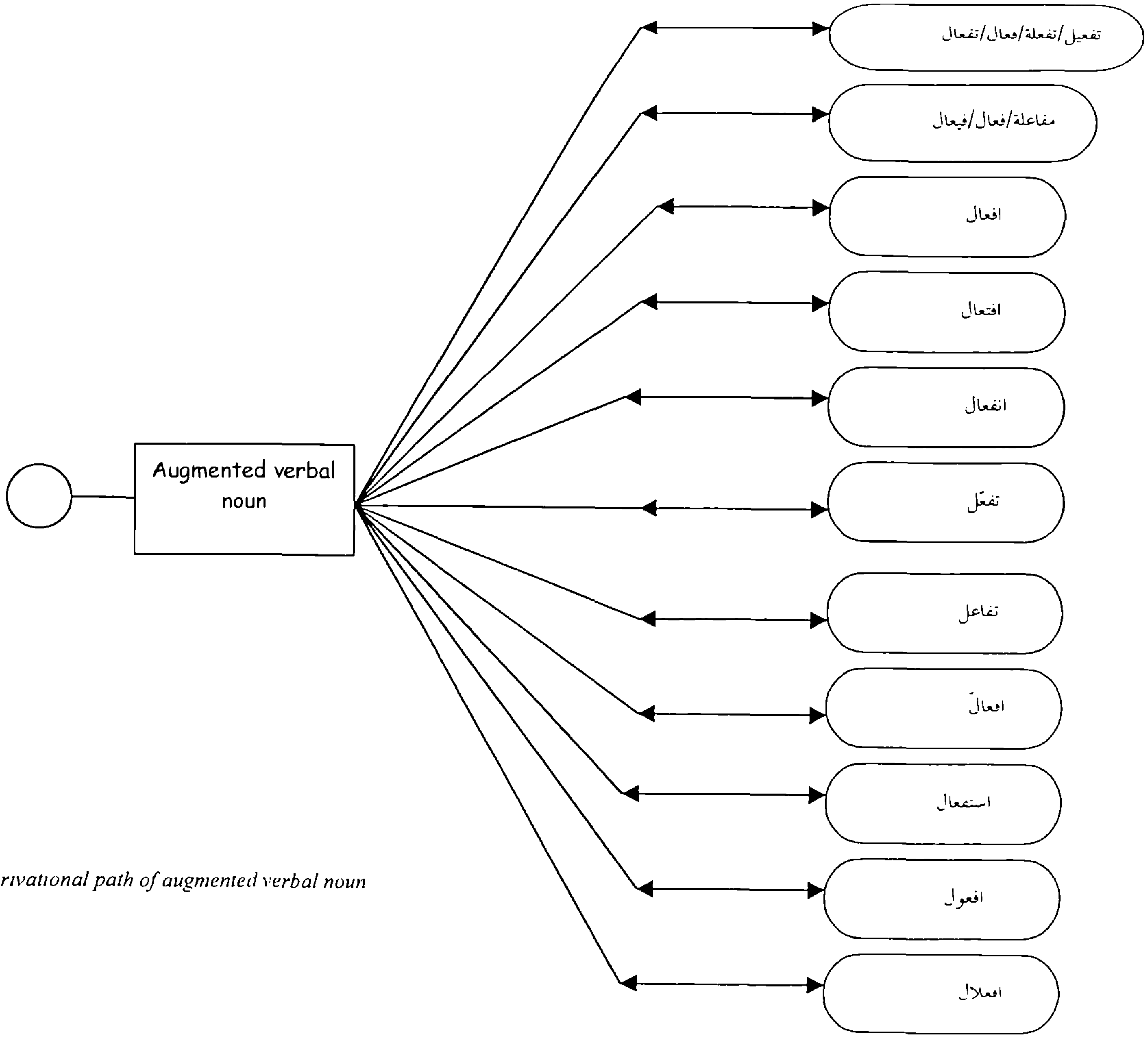


Figure 6.7 Derivational path of augmented verbal noun

The derivational path of the two forms can be represented as shown below (Figure 6 8)

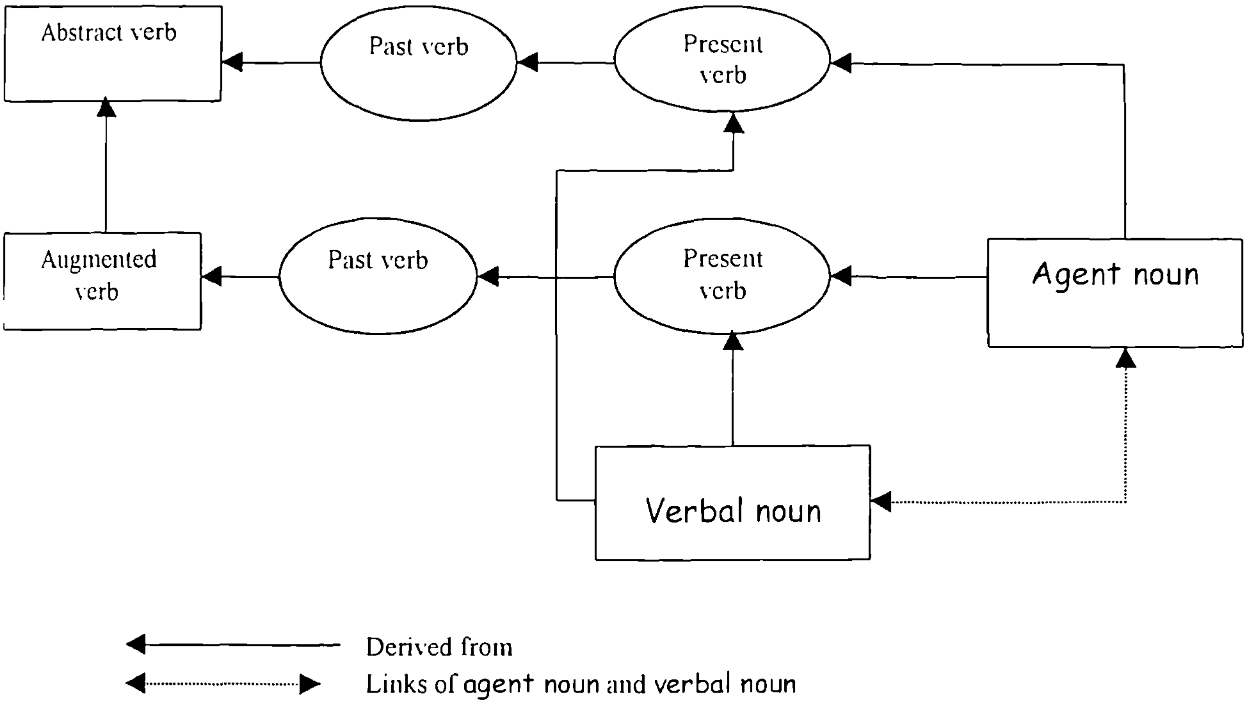


Figure 6.8 An example of a semantic link between the agent noun and the verbal noun forms

6.5 Inflectional morphology

It was mentioned above that the inflectional affixes are added when the whole derivational process is completed. It is important here to differentiate between derivational and inflectional morphology in Arabic. The function of the inflection is to alter the form of the word in number, gender, mood, tense, aspect, person, and case (Klavans and Tzoukermann, 1992) while derivation may change the grammatical category of a word. There are a number of specific differences between derivation and inflection which may be summarised as follows (Ali, 1988):

- Inflection deals with syntactically determined affixation processes while derivational morphology is used to create new lexical items;
- Inflection is regular, while derivation is not;
- Inflection affixes are placed on the terminal point of the word, while the derivation affects the structure of the word;
- Inflectional morphology has a strong relationship with syntax.

6.5.1 Inflectional affixes (prefixes and suffixes)

The inflectional affixes can be defined as bound grammatical morphemes that are added to complete words according to rules of syntax, (e.g. *third person singular verbal suffix -s*). The inflectional affixes can be divided into prefixes and suffixes. Prefixes come before the form to which they are joined, such as the definite article and preposition prefixes. Suffixes come after the form to which they are joined, such as gender, number, tense, and person. Table 6.4 is an example of some suffixes joined to the root *KTB* (to write)

Table 6.4 some suffix inflections for the root KTB

	Word in Arabic	Transliteration	Suffixes	Meaning
Singular	كتب	katab	a	he wrote
	كتبت	katab	at	she wrote
	كتبت	katab	ta	you wrote(m)
	كتبتي	katab	ti	you wrote(f)
	كتبت	katab	tu	I wrote(c)
Dual	كتبا	katab	a	they(two)wrote (m)
	كتبتا	katab	ata	they(two)wrote (f)
	كتبتما	katab	tuma	you(two)wrote (c)
Plural	كتبوا	katab	u	they wrote (m)
	كتبن	katab	na	they wrote (f)
	كتبتم	katab	tum	you wrote (m)
	كتبتن	katab	tunna	you wrote (f)
	كتبنا	katab	na	we wrote (c)

6.5.2 Inflection Case (parsing change)

Syntax in Arabic also has an impact on word formation. The syntax can be defined as the study of grammatical relations between words and other units within the sentence (Matthews, 1997). The meaning of syntax in Arabic is the changing which takes place to the word ending according to the part of speech (tools) which precede the word. For example, the suffix of

المدرســـون *almuDaReeSun* (teachers) will change according to the previous tools, as shown below:

1. جاء المدرسون *Jaa almuDaReeSun* (The teachers came).
2. رأيت المدرسين *shahatu almuDaReeSin* (I saw the teachers).
3. ذهبت مع المدرسين *dabatu maa almuDaReeSin* (I went with the teachers).

It is noted that the suffix المدرسون changed because the word (teachers) in the first example is فاعل *fa'al* (the subject), while it is the مفعول به *maful* (the object) in the second example. In the third example, the word is اسم مجرور *Ism majurror* (noun in the genitive) and the preposition is مع *ma'a* (with).

In Arabic the case inflections can be categorised as follows:

1- case inflections of noun:

- □ Case 1 (nominative case);
- □ Case 2 (accusative case);
- □ Case 3 (genitive case);

2- case inflection of verb:

- □ Case 1 (nominative case);
- □ Case 2 (accusative case);
- □ Case 3 (apocopative case).

6.5.3 Combining form

In Arabic, it is common to combine a word with tools, such as prepositions and conjunctions. Nouns and verbs can be attached to a number of prefixes and suffixes. Combining forms causes numerous problems in Arabic. For example, in automatic morphological analysis, the system must first distinguish between syntax and morphology before morphological analysis can start.

6.6 Morpho-phonological changing

The most important aspect of morpho-phonological changing is the mutation and vocalisation cases. The mutation or morphological substitution can be defined as the changing or removing of a consonant letter and its replacement with another. For example, to derive the word اصطبار (which means: *long-suffering*) from the root صبر (which means: *to be patient*), the following morpho-phonological changing should take place:

- The equivalent morphological form of the word اصطبار is افتعال *iFtiAaL*,
- The third letter in the morphological form is ت (*ta*), while the third letter in the word is ط (*ta*). Though these are pronounced the same in English, they are in fact different letters,
- Morphologically, the third letter should be changed from ت (which is a part of the morphological form) to ط (which is the correct letter that should take a place in the word). The morpho-phonological change can be represented as follows (Figure 6.9):

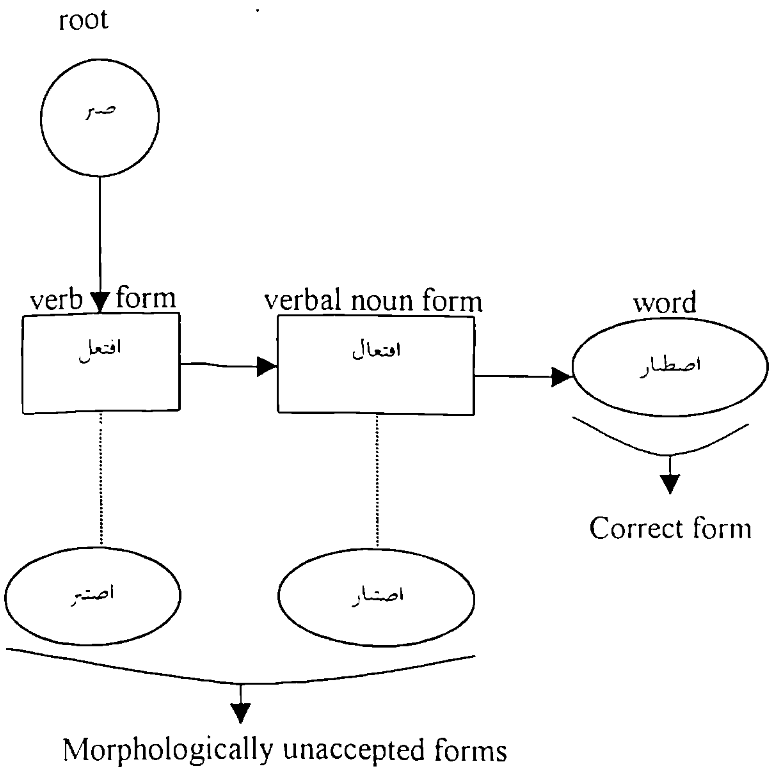


Figure 6.9 morpho-phonological change of an Arabic word

Vocalisation is a part of phonology which deals with a word that has weak letters within its structure (weak letters are: ا as *a* in English , و as *w* in English, and ي as *y* in English). The vocalisation process is to change the weak letter or to delete it. For example, the word سماء (sky) can be reduced to its root by changing the terminal letter from (ء) to (و) thus

سماء ⟶ سمو

6.7 Summary

This chapter has covered some concepts of the morphological system in Arabic. The morphological system in Arabic plays a major role in word formation. This in turn will affect the retrieval performance in Arabic. Morphology in Arabic can be divided into two types: derivation and inflection. Each of these methods can be further divided into types. The two types of Arabic morphology were discussed in detail. Both types have an impact on word formation in Arabic, but derivational morphology has more influence than inflectional morphology.

Derivation-based Morphological forms is the most important and recognisable type. The majority of Arabic words are based on this type of derivation. Each morphological form consists of a root and a pattern (composed form). In order to generate an Arabic word from the root, one should follow the morphological forms. This can be done via derivational path which consists of several rules such as rules of derivation, rules of morphological forms and phonological changing. The function of the morphological rules is to control the derivational process through the derivational path.

An effort has been made to represent the Arabic derivational path in a way that can be understood by a reader who is not familiar with the language. The Arabic root system, in addition to the morphological form rules were investigated. This chapter represents the conceptual framework of the study. The implementation of the framework is discussed in the next chapter.

Chapter Seven: System architecture and implementation

7.1 Introduction

This chapter deals with the prototype components such as inverted files, search methods and the knowledge base, in addition to the morphological processor. The prototype was developed to implement the morphological theories, which were mentioned in Chapter 6. This chapter starts with an overview of the system, followed by its components. The problems encountered during the development of the system are discussed. The morpho-semantic approach, which is introduced by this study, is discussed in greater detail. The chapter ends with a description of the morphological processor and its module, including the functionality of each module.

7.2 Overview of the system

A prototype of an Arabic information retrieval system was developed using Prolog (Amzi! Prolog Version 4). Figure 7.1 depicts the main building blocks of the system. Later in this chapter, further discussion for each block is given, but before that, it is important to give an overview of the system. For the study, a collection of titles, abstracts, and full texts written in Arabic were used as a database. This database can be accessed through a number of methods: word, stem, root, and morpho-semantic. The prototype was supported by a morphological analyser to carry out the morphological analysis tasks. The main aim of developing the current system is to investigate the retrieval performance of the novel morpho-semantic method against other methods used in Arabic information retrieval systems (i.e. word, stem, and root). The morpho-semantic method is based on the idea that linking some morphological forms sharing some semantic features can improve retrieval performance. This claim is discussed in Chapter 8.

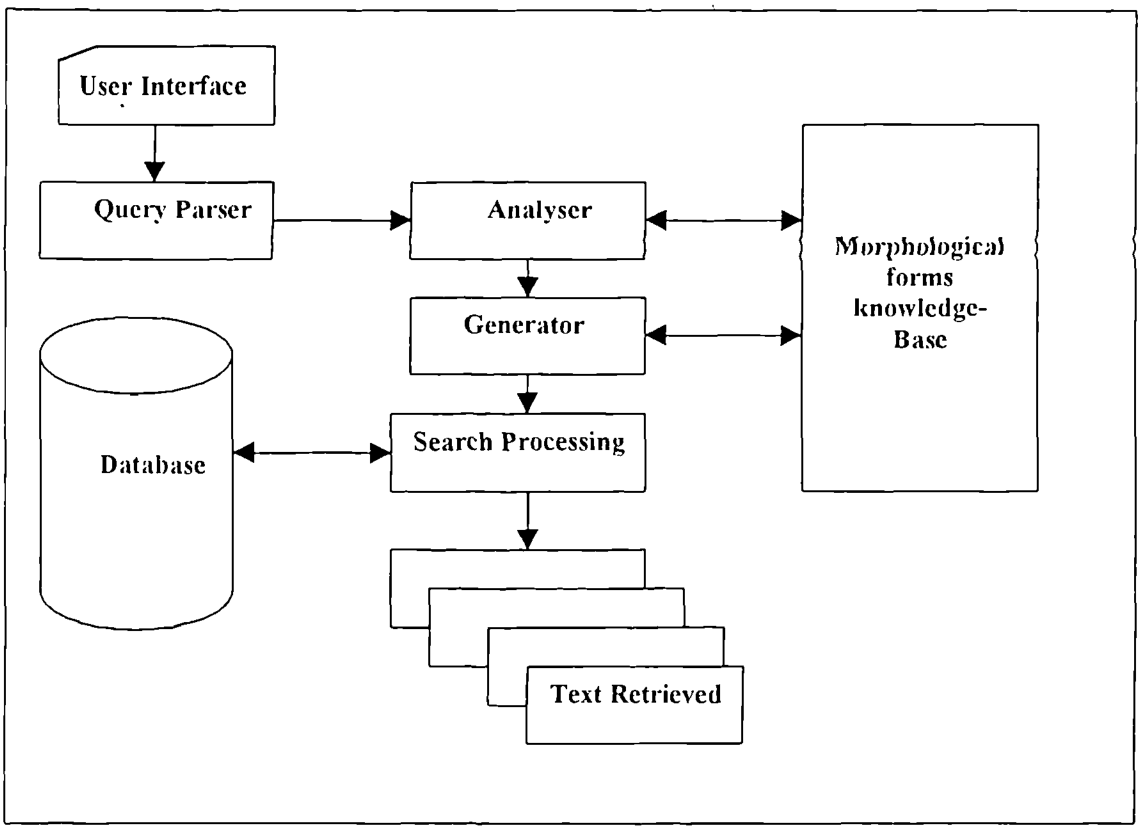


Figure 7.1 Outline Architecture of the Text Retrieval System (AIRSMA)

7.3 Interface

Developing a sophisticated interface for the current system is beyond the scope of the study (at least for the present time). Therefore, a simple menu driven interface was developed to allow the user to interact with the system. When the user runs the system, the following menu appears:

```
*******************************
*        اختر منهج البحث        *
*******************************
*   ١            منهج الكلمة      *
*   ٢        منهج ساق الكلمة      *
*   ٣        منهج جذر الكلمة      *
*   ٤         منهج صرف دلالي      *
*   ٥                   خروج      *
*******************************
```

Arabic interface

```
*******************************
*     Choose search method    *
*******************************
*  1. Word method             *
*  2. Stem method             *
*  3. Root method             *
*  4. Morpho-semantic method  *
*  5. Exit                    *
*******************************
```

English interface

When a user chooses a search method, the system asks the user to enter his/her queries. The system also expects single or phrase terms. Users can also use Boolean operators, especially AND and OR. When a query is input into the system, it will be treated according to the search method mode. For example, if the user chooses a word method, the system will search the inverted file of the word terms without doing any thing to the user query, rather than matching the query and terms related to it in the inverted file of word terms. However, if the user chooses the stem method, the system will treat the query differently. In other words, the system parses the query by removing any prefixes or suffixes that may be attached to the given query. The same thing applies to the remaining methods (i.e. root and morpho-semantic). A further description of each method is given below.

7.4 Database description

In Chapter 5, a detailed description of the sample collection (database records) was given. This section concentrates on how the database records are represented in Prolog. After reviewing the records (the sample of study), all the 590 records were divided into 5 groups, each of which has a different data representation. All five groups were represented in the following schema:

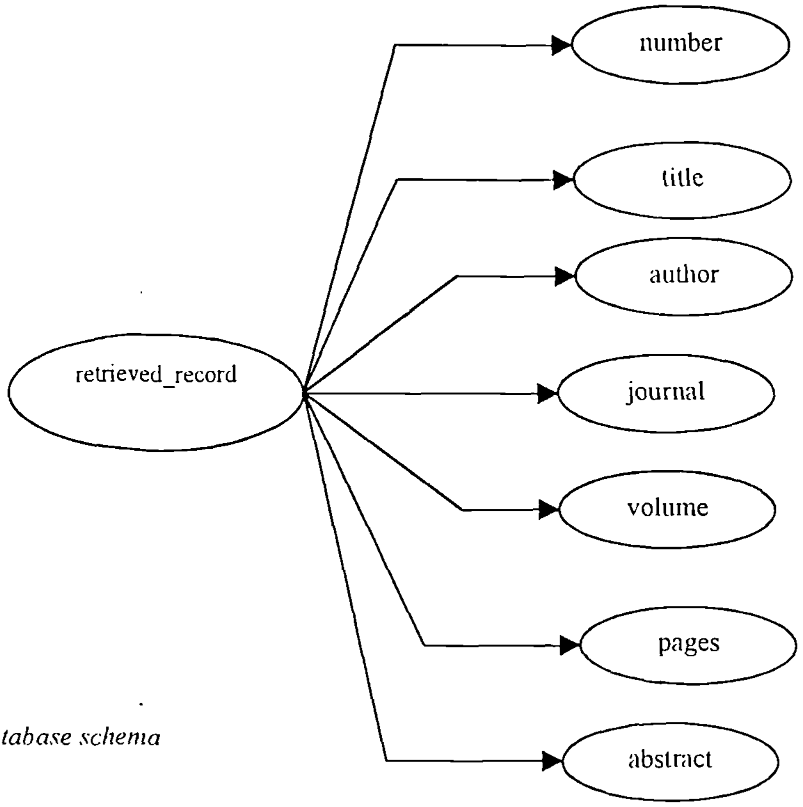


Figure 7.2 database schema

However, not all the fields are present in all records. For example, in group number five about economics the only fields present are: number, author, title, journal, pages. The above relationship schema was represented in Prolog predicates as follows:

```
retrieve_record(Number,Title,Author,Journal,Volume,Pages,Abstract):-
            record(
            number(Number),
            title(Title),
            author(Author),
            journal(Journal),
            volume(Volume),
            pages(Pages)
            abstract(Abstract)).
```

Figure 7.3 Prolog rule to retrieve the record (document) entity

Example of a database record and its Prolog rule representation is given in Appendix 2

7.5 Index (Inverted) files

The organization in the inverted file is turned around (inverted) to create an index for all unique key values in all documents. A file in which the items themselves provide the main order of the file is known as a direct file. The inverted index, on the other hand, is arranged in order by topic, and each topic includes the corresponding list of item numbers. When an inverted index is available, each topic term is then usable as a key to obtain access to the corresponding items.

The inverted file ensures quick access to the information items because the index alone is examined in order to determine the items which satisfy the search request, rather than the actual file items (Salton and McGill, 1987). A number of inverted files were created. The following is a brief discussion of how the inverted files of the system were developed.

7.5.1 Word index file

This file was created from the plain texts (database records). All the 590 records were converted to inverted files (individual index terms). To implement this, the following steps were taken:

- All words of the text documents must be identified;
- The stop list words should be removed from the index files;
- Any duplications should be removed,
- Index terms and their assignment to the documents of collection should be identified.

To carry out the above task, a prolog program was developed to convert the plain texts into accessible index terms. The output of the program is a list of index terms. These index terms were represented by Prolog as the following facts:

```
word_term('word',pointer).
```

Word term indicates the predicate of the index terms for words. This predicate has two arguments: the first argument is the index term for word, while the second argument is a pointer (record number). This pointer is used to access the real documents where the index terms occur on them.

Two fields are used to generate the inverted file for the index word. These fields are: title articles, in some cases, and abstracts in others. The inverted file of word index terms is used to support the search using the word method. All other inverted files of the system (i.e. stem and root files) are based on the word index file.

7.5.2 Stem index file

The stem index is derived from the word index file. All the index terms that appear in the word index file are converted to stem terms, and stored in the stem index file. To do so, all prefixes and suffixes that may be attached to index (word) terms should be extracted (i.e reducing the original words to word stems). This can be achieved by using the prefix and suffix removal model (part of the morphological processor). Reducing word index terms to their stems will result in some duplication. These duplications need to be removed from the stem index file. The stem index terms were represented in the Prolog database as a predicate with two arguments as the following fact:

```
stem_term('stem',pointer).
```

Reducing a word to its stem may sometimes create some difficulty for information retrieval. Below are examples of some of the difficulties encountered whilst developing the stem index file.

First example:

It is known, and this was mentioned above in Chapters 3 and 6, that some particles or prefixes and suffixes in Arabic are joined with other words. This joining creates some difficulty for information retrieval. One way of dealing with these particles or prefixes and suffixes is to remove them. However, some of these prefixes and suffixes may appear in some words as a radical character while in others they may appear as part of a word. For example, the character و *wa* is joined with the following word as a prefix thus:

واسترجاع (*wasterja'a*), while the same character و *wa* is a radical character in the following word: وسائل (*wasail*).

To solve this problem (i.e. removing prefixes or suffixes), there are two approaches: automatic and manual. In this work a hybrid approach (combining both automatic and manual) is used. An automatic approach was used based on the following algorithms:

- □ scan the word and break it into a list,
- □ check the lists of prefixes and suffixes,
- □ if a match is found in the list, remove the matched characters from the given word,
- □ check the morphological forms' database to find equivalent forms for the given word,
- □ if a match is found, put the word in the stem file as an index term,
- □ else do not remove any character from the given word.

However, some prefixes and suffixes were removed manually. In some cases these were prefixes or suffixes, and in other cases they were radical, as in the example given above.

Second example:

In this example, the problem is related to dual and sound plural suffixes. It was mentioned in Chapter 3 that there are a number of suffixes for dual and sound plurals. These suffixes should be removed in order to unify all the forms of the number (singular, dual and plural) under one entry (i.e. singular form). This is not as easy as it is in English, which requires the removal of the *-s* only in order to generate the singular. (There are some exceptions, but the *-s* ending is the one mostly used for plural in English). In Arabic, the situation is different as previously mentioned; there is a single type of suffix for dual and plural. It would be pointless to list everything that was done to solve prefixes and suffixes, but one example may be enough.

The example which will be mentioned here is the suffix ات which is used to derive the feminine sound plural. Suppose that the system removes this suffix from a given word without any rules or conditions. This would create a problem with words where the same characters (ات) appear at the end of the word but are radical parts of the word rather than suffixes. For example, in a word such as مات (meaning 'died') removing 'ات' would lead to a nonsensical term 'م'. In order to overcome this problem, the following rule is used:

RULE1:

IF a given word ends with suffix ات AND
The word has more than three characters,
THEN remove the ات .

This rule will not remove the suffix ات if the given word has fewer than four characters. However, the issue is not constrained by the number of characters in the word and more rules are needed to the solve problems of the suffix ات . The same suffix is sometimes used as a suffix of several morphological forms (e.g. **noun** or **verbal noun**).

□ **noun** as in دراجات which means bikes

□ **verbal noun** as in تشريعات which means legislations

Both words end with the same suffix ات . This suffix is a sign of feminine sound plural but the morphological form for each one is different To remove the suffix to reduce the word to its single form, it is correct to remove the suffix from the **verbal noun** without doing anything except remove the suffix to get the singular form (تشـــريع), which is equivalent to the morphological form (تفعيل). However, it is not appropriate to do the same thing with the **noun** form, the result would be (دراج), while the correct word is (دراجة), which is equivalent to the morphological form (فعالة). The correct form of this word can be achieved by removing the suffix ات and adding the suffix ة which indicates the singular of the feminine. To do so the following rules are applied:

RULE2:

IF the given word ends with suffix ات AND
The word has more than three characters, AND,
The word is not present in the X list,
THEN remove the suffix ات and add the suffix ة .

RULE3:

IF a given word ends with suffix ات AND
The word has more than three characters, AND,
The word is present in the X list,
THEN remove the suffix ات .

Where the X list contains some morphological forms which were created to solve the ambiguity of the suffix ات.

7.5.3 Root index

In order to create a root file, a full list of Arabic morphological forms is needed, in addition to the morphological analyser which is capable of reducing a given word or a stem to its root. Further details of morphological forms and the analyser will be discussed later in this chapter. The stem file will be used to generate the root terms file. All stems sharing the same root are reduced to a single root. For example, the following stems: تدريس (teaching), مدرس (teacher), مدارس (schools), and دراسة (studying) are reduced to the single root درس (to teach). The root index was designed to locate all possible word derivative forms of a single root. During the creation of the root file, several problems were faced related to Arabic morphology such as weak forms and morpho-phonological rules. (See Section 6.1.2.4 as an example.) However, covering all Arabic morphological subjects is beyond the scope of the current study. Further details of the Arabic morphological knowledge base are covered in the following sections.

7.6 Search processing (methods)

It was mentioned above that users can choose one of four methods (word, stem, root, or morpho-semantic) to search the database, each of which has its advantages and disadvantages. However, a full comparative study of performance of each method is discussed in Chapter 8 (system evaluation). In this section a brief description is given of how these methods are working. It should be mentioned here that the first three methods (i.e. word, stem, and root) are available on some commercial retrieval systems in Arabic and they have been investigated by some Arab information evaluators, as was discussed in Chapter 4. The fourth method is a novel approach and is introduced in this study which is believed to improve information retrieval in Arabic. The following sections discuss each method.

7.6.1 Word method

This method deals with a word as it appears whether in a document or a query. In other words, if this method is used by the user, the system will search for an exact match between the user queries and the keywords which appear in the document. This will be achieved by consulting the inverted file of the word terms. After the successful matching is achieved, the system retrieves the items related to the user query. Unfortunately, words have many morphological variants which will not be recognised by exact-matching algorithms without some form of natural language processing. In most cases, these variants have similar semantic interpretations and can be treated as equivalents for information retrieval applications (Hull, 1996). Because of the absence of linguistic analysis for the word method, this method may miss some relevant documents.

7.6.2 Stem method

This method is similar to the word method with some linguistic treatment for both documents and queries. This method is based on the word stem rather than the word itself. In other words, all prefixes and suffixes that may be attached to words are removed in order to unify words having the same stem under the same stem. This may increase the success of matching documents to a query. As is the case with the word method, this method may miss some relevant documents.

7.6.3 Root method

Because Arabic is a derivative language, this has led some information retrieval system developers to design retrieval systems based on the root method. The idea of this method is that all words having the same root should be reduced to a single root. In other words, this method is able to retrieve all the morphological variants of a query or keywords. However, this method is criticized by some information specialists and some librarians (as was mentioned in Chapter 4) for retrieving unwanted documents.

The existing Arabic text retrieval system uses the above three methods. Each of these methods has its limitations. The word and stem methods may miss relevant texts, while the root method may retrieve irrelevant texts. This study uses these methods to compare the performances of each and against the newly proposed method, which is discussed below.

7.6.4 Morpho-semantic method

Having said that each method has its limitations, this motivates the investigation of a different approach for Arabic information retrieval. The proposed new approach is called here the morpho-semantic method. This method is based on the idea that creating links between some morphological forms (especially those forms sharing close meaning such verbal noun and agent noun) may improve retrieval performance in Arabic. It would be easy for both the writer and the reader to explain this method by giving the following example, which compares the retrieval performance of the morpho-semantic method against other methods (word, stem and root).

For readability purposes, let us consider the root شـــرق *SHRQ*, (the general meaning of this root is something related to "east") and see how words can be generated from this single root via the semantic networks shown in Figure 7.4. If the process starts from the root (node 1), all possible forms would be generated. The following represent some meanings of the generated words:

- word in position 13 شروق (*ShuRoQ*) means "sunrise";
- word in position 11 مشرق (*maSHReQ*) means "the place of the sunrise";
- word in position 12 مشـــارق (*maSHaReQ*) is the plural of the word in position 11;
- word in position 4 مستشرق (*mostaSHReQ*) means "orientalist";
- word in position 5 استشراق (*isteSHRaQ*) mean "orientalism";
- word in position 8 تشريق means "the days following the day of sacrifice".

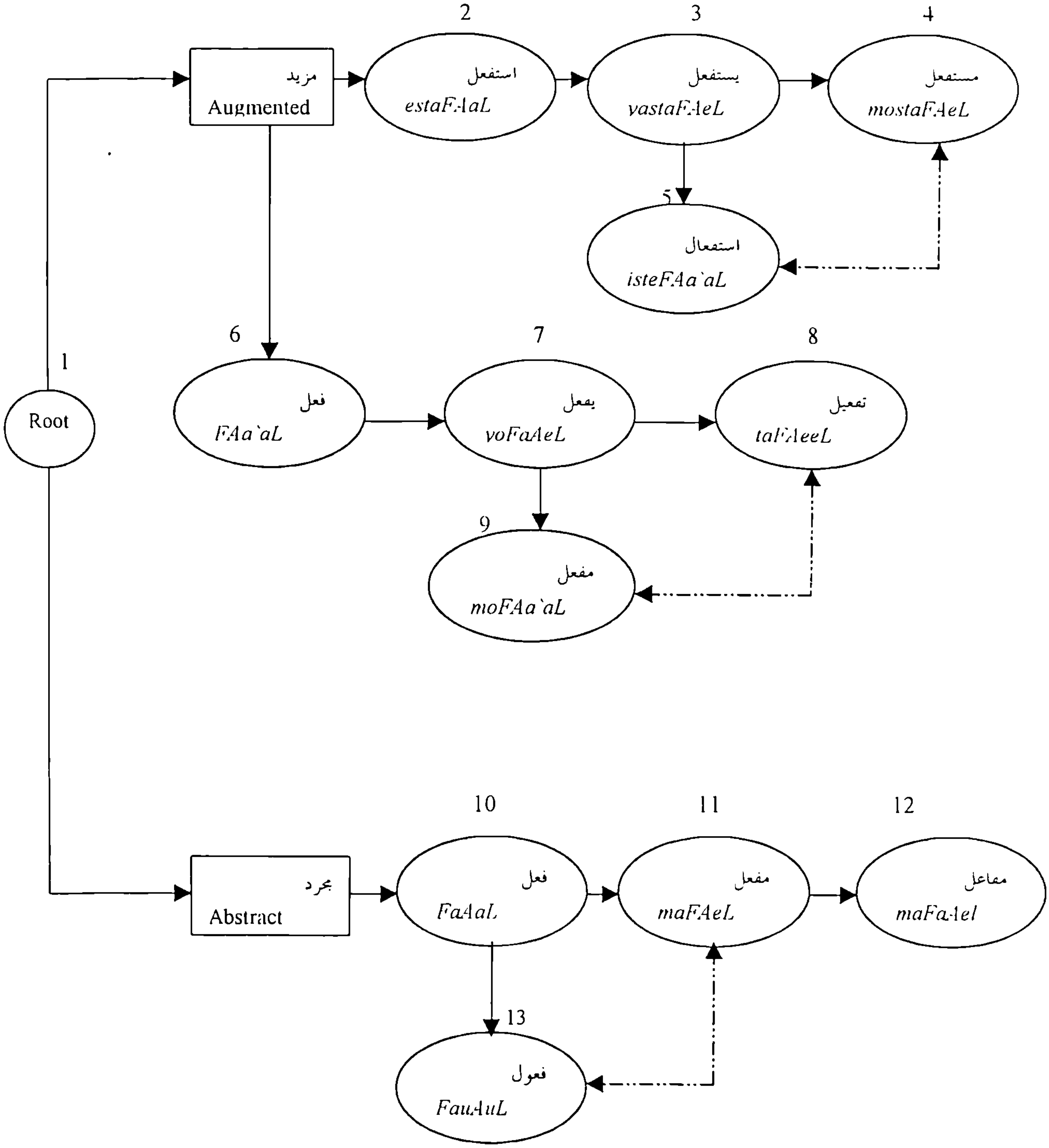


Figure 7.4 an example of semantic networks for morphological forms

Suppose a user wants to search for the term in position 4 (مستشرق `*mostaSHReQ*', orientalist). If he or she uses the root method to retrieve related terms to his or her query, all the terms in positions 1-13 would be retrieved. This means that the wanted terms (in our case terms in positions 2, 3, 4, and 5) and unwanted ones would be retrieved. However, if the word method is used, the only term that would be retrieved is the term in position 4. This means that the word method would miss the related terms in positions 2, 3, and 5, in addition to any terms attached by any suffixes or prefixes in those positions.

When the stem method is used, the term in position 4 would be retrieved, in addition to all terms attached by any suffixes or prefixes in that position. If the user uses the morpho-semantic form method, the system would retrieve terms in positions 2, 3, 4, and 5, in addition to all term forms in those positions. This is due to the semantic link that has been made between the forms (as shown by the dotted line), and the internal relationship between the forms which is represented by the arrow ——▶ as shown in Figure 7.4. It has been noted that the general meaning of the root is wider at the first position (i.e. 1). The more the user moves forward the more specific the meaning becomes, as shown in positions 2 and 3. This semantic representation is applied to all the related morphological forms.

This example shows how the morpho-semantic approach promises to improve the effectiveness of the word, stem and root methods.

7.7 Knowledge base

7.7.1 Prefix and suffix lists

In order to analyse any Arabic word, it is important to identify or to distinguish between the radical characters of the word and the augmented ones. To do so, the prefixes and suffixes that may be attached to the Arabic words should be identified in advance. Therefore, most of the prefixes and suffixes used in Arabic were collected. Then, two lists were created. The first list is for prefixes and the other one for suffixes. Different ways were examined to represent them. Finally the following representation was reached. Each prefix or suffix was put in a list (list in Prolog is an important key for problem solving). Each of them was represented as facts (fact in Prolog consists of a predicate, and zero, one or more arguments. If a fact is present this indicates it is true, if not it is not true). Thus:

```
preffix_list([ال]).
suffix_list([ون]).
```

The above predicate means that, since ال is present as a fact in the prefix list, the ال is considered as a prefix. Therefore, if this prefix appears in a word it should be removed. This same rule can be applied to suffixes. The two lists support the prefix and suffix removal module to complete its task. In order to remove a prefix or suffix from a word the following predicates are used:

```
remove_prefix(Word,Stem).
remove_suffix(Word,Stem).
```

To remove both prefix and suffix from a given word, the following predicate was used:

```
stem(Word,Stem).
```

7.7.2 Morphological forms representation

As mentioned above, morphological forms are at the heart of Arabic morphology. In order to extract a root from a given word, the equivalent morphological form of the given word must be known. For example, the word دراسة **DeRaSah** (studying) can be segmented thus:

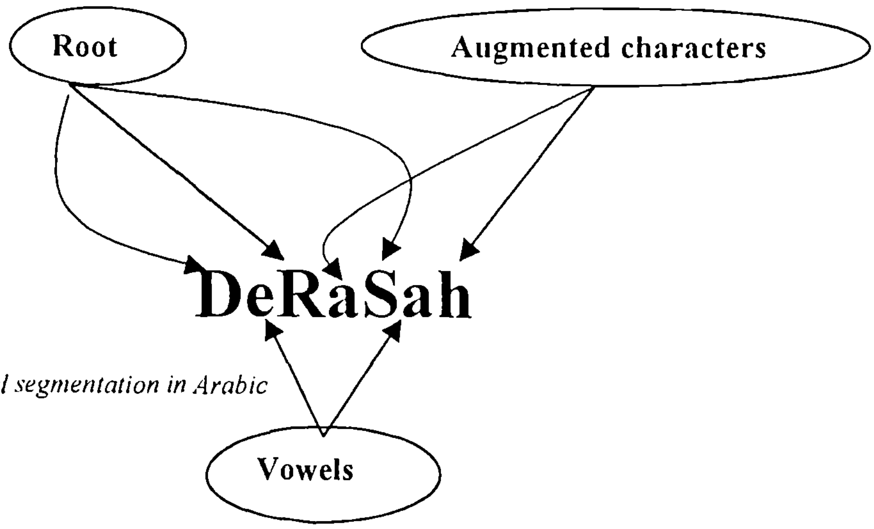
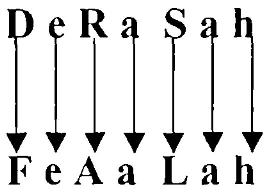


Figure 7.5 word segmentation in Arabic

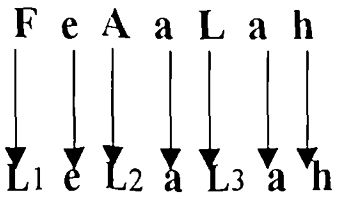
The equivalent morphological form of the above word is فعالة **FeAaLah.** Capital characters indicate the radical characters of the word (root). If a match between the above word and the equivalent morphological form, is found in the morphological form knowledge base, the match would look like the following:



The above is a brief description of how the morphological forms work. In this work most of the morphological forms used in Arabic were collected. These forms were collected from various sources such as books of morphology in Arabic, tables, and dictionaries. (See Appendix 1.) Once the collection is completed, each morphological form is classified according to its character length. It was found that there are seven classes of lengths that can be represented as follows:

- □ **weight_2** for the bilateral-lettered length
- □ **weight_3** for the third-lettered length
- □ **weight_4** for the fourth-lettered length
- □ **weight_5** for the fifth-lettered length
- □ **weight_6** for the sixth-lettered length
- □ **weight_7** for the seventh-lettered length
- □ **weight_8** for the eighth-lettered length

Each morphological form takes one of the above classes, and is represented as a list of variables and constant characters. Thus:



For example, the morphological form **FeAaLah** can be represented in the morphological form knowledge base as the following fact:

weight_7([L1,e,L2,a,L3,a,h]).

The variables **L1**, **L2**, and **L3** represent the radical characters (root) of the morphological form; and the remaining characters are the constants. Such a representation can be used to analyse a given word to its root, and at the same time, it can be used to generate several words from a single root.

The main aim of developing the morphological forms is to enable the system to deal with changes that occur in Arabic words. Furthermore, through using morphological forms, it becomes easy to recognize (or extract) the word root. This is explained through the following example. Consider representing the fact that “تدريس” is reduced_to the root درس “, in English “*teaching* is reduced_to *teach*”. This fact consists of two objects, called “تـــدريس” (morphological form) and “درس’ (root), and a relationship called “reduced_to”. In Prolog, facts must be written in a standard form, like this (Clocksin and Mellish, 1994):

reduced_to("تدريس","درس").

The interpretation in English for the above fact is that the word “تدريس” is reduced to the root “درس”. In other words, the root of “تدريس” is “درس”. The morphological forms in the knowledge base are represented as a skeleton. For example, the word تدريس can be represented by the morphological form تفعـيل . Prolog is used to represent the morphological form and its root as follows:

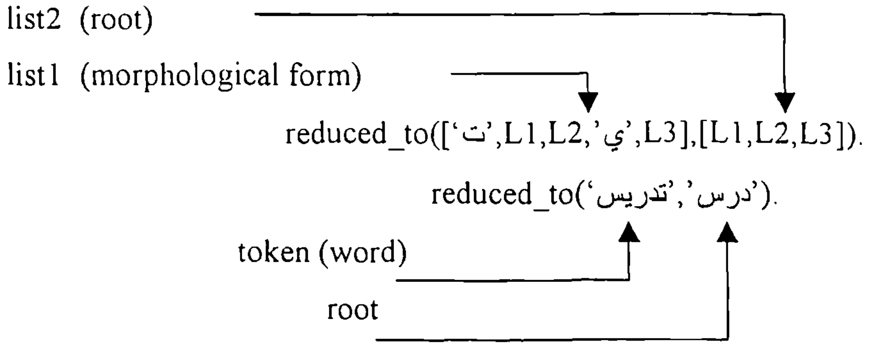

list2 (root)
list1 (morphological form)
reduced_to(['ت',L1,L2,'ي',L3],[L1,L2,L3]).
reduced_to('تدريس','درس').
token (word)
root

Figure 7.6 morphological forms representation

The first list stands for the morphological form, while the second list stands for its root. Characters in positions **L1**, **L2**, and **L3** are variables and represent the radical characters (root) of the morphological form; and the remaining characters are the constants. To reduce the morphological form (list1) to its root (list2) constants should be removed. The above morphological form (list1) can be applied to hundreds of words in Arabic such as تمريض, تعليم, تخزين, تحنيط, تنويم, تقسيم, تفسير and so on.

7.7.3 Semantic networks for morpho-semantic forms

In the previous section the morphological forms representation was discussed In fact, the above representation is supporting only the root method. The main disadvantage of this representation is that the irrelevant words (sharing the same root) will be reduced to a single root. For example, the words جامع (mosque), جامعة (university), مجموعة (group), and اجتماع (meeting), all these words in Arabic has the same root جمع (to gather). Generally speaking, it is true that all the above words share in some way the meaning of the root (to gather). For example, people gathering in a mosque, at a university, in a meeting or as a group.

For information retrieval, this type of analysis may not be required (i.e. it is necessary to reduce some forms of word to its root, but at the same time unrelated ones are not wanted). To some extent, reducing word forms to a single root is useful for Arabic in some, but not all, situations. Generally speaking, the root method, as reported by previous studies Al Tayyar and Bechkoum (1998); Al kharashi and Evans (1995) can be useful if a high recall is required but without high precision. However, the word and stem methods show that they are effective when high precision is needed. This researcher believes that an optimum solution is somewhere in between and has therefore given considerable thought for solving the above problem. A novel method was developed and given the name the morpho-semantic method. This method has the advantages of both the stem and the root method. It is based on linking morphological forms that are related to each other, such as forms of number in Arabic or verbal nouns and agent nouns. To implement this method, a number of semantic links between morphological forms have been developed and are discussed below.

7.7.3.1 Internal and external links

Two semantic links are developed to link between derivatives: external and internal links External links are made between types of morphological derivatives. For example, links are made between verbal nouns and agent nouns. Within these two types, further links can be made. In other words, within the verbal noun, there are several forms related to each other that can be linked to each other. (See Table 7.1.) In addition to this, within some derivatives, a link can be made between classes as in forms of number (see next section), which contain singular, dual, and plural. Figure 7.7 shows internal and external links between derivatives and morphological forms.

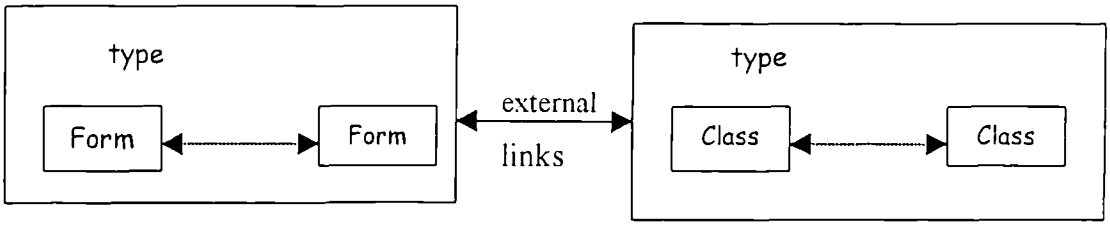


Figure 7.7 internal and external semantic links for morphological forms

I. Number form links

There is a strong relationship between forms that are singular, dual, or plural (whether broken or sound). These types of forms are used in Arabic to express concepts such as objects and persons or an abstract concept. For example, the words شجرة (one tree as a singular), شجرتان (two trees as a dual), شجرات (trees as a sound plural for feminine) and أشجار or شجر (trees as a broken plural) are used to express an object in different forms. Since these forms are related to a single object, it would be useful to link them to each other For implementation purposes, a semantic network was developed to represent all the relationships of number forms, as shown in Figure 7.8.

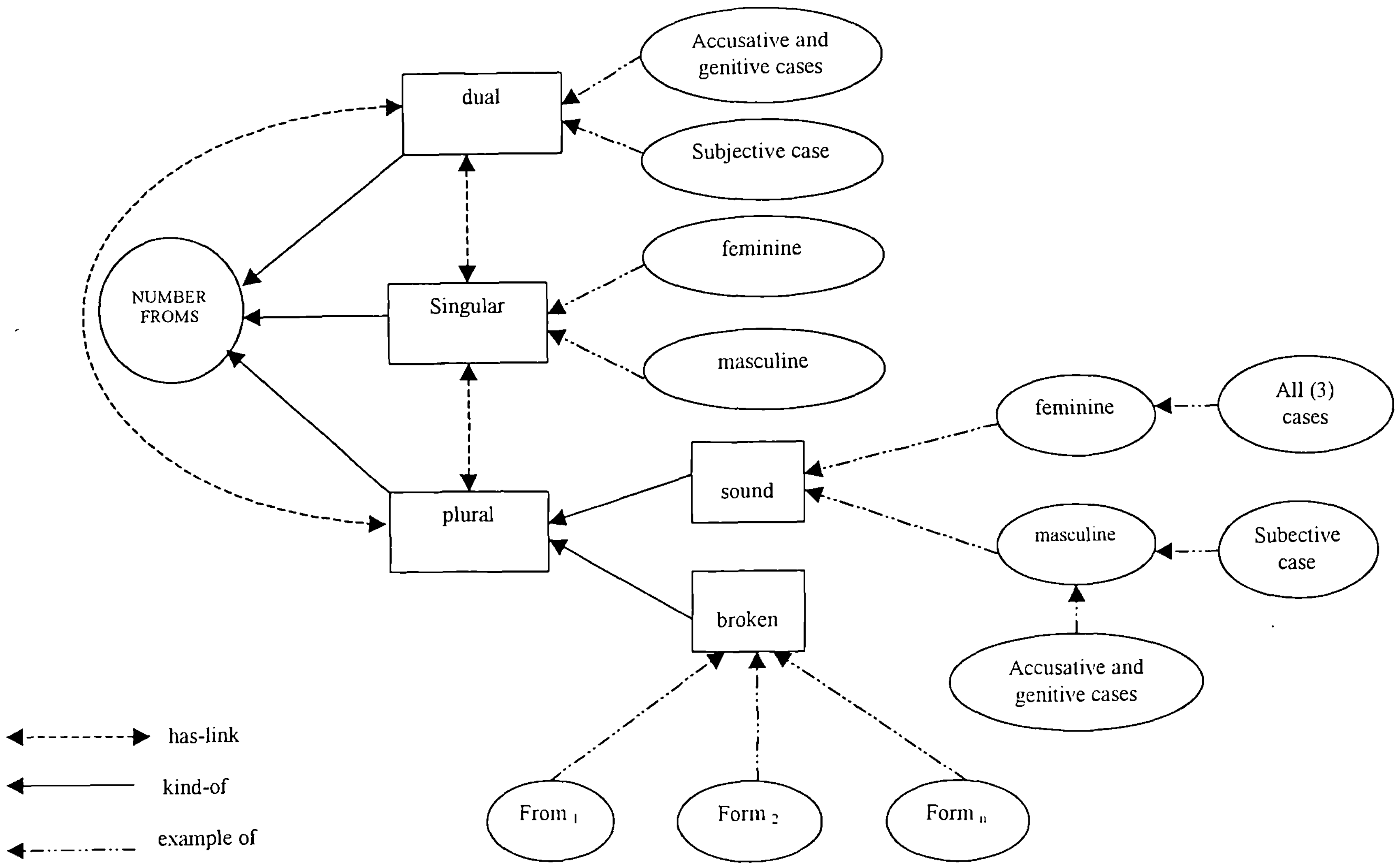
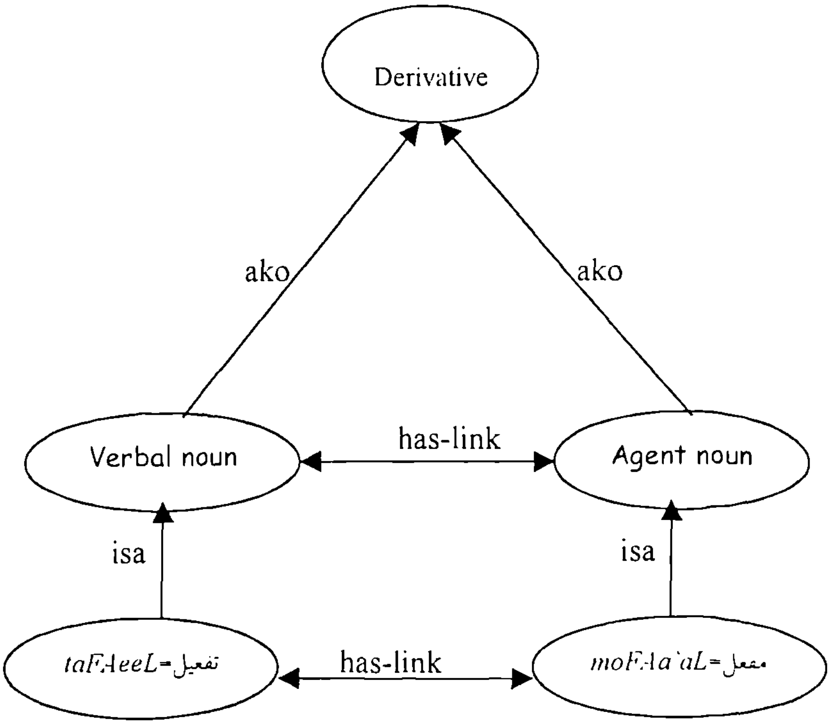


Figure 7.8 semantic networks for NUMBER forms in Arabic

II. Verbal noun and agent noun links

This is another kind of semantic link implemented in the study. Both verbal noun classes and agent noun classes are words of other classes derived from verbs. For example, suppose one wants to derive the verbal noun from the verb *act*, the suffix *-ion* should be added to the verb in order to get the word *action*. However, if the agent noun is required, the suffix *-or* should be added to the verb *act* to get *actor*. Since both forms are derived from the same verb, a semantic link can be generated. In Arabic, there are several morphological forms that can be treated in a similar way. Figure 7.9 shows the semantic networks of verbal noun and agent noun.



ako stand for 'a kind of', meaning that something is one of several kinds of something else
isa stand for 'is a', indicating that something is a named example of something else

Figure 7.9 example of semantic links for verbal noun and agent noun

From the above semantic networks, hundreds of words can be generated. Table 7 1 shows a few examples of words linked to each other.

Table 7.1 examples of words based on semantic links

Agent noun	Verbal noun
ممرّض (nurse)	تمريض (nursing)
مدرّس (teacher)	تدريس (teaching)
موجّه (instructor)	توجيه (instruction)
مورّد (importer)	توريد (importing)
مولّد (generator)	توليد (generating)
ممرّن (trainer)	تمرين (training)
مشرّع (legislator)	تشريع (legislation)

Furthermore, one concept may be represented in different morphological forms. For example, the following verbal noun forms فعال and مفاعلة give the same meaning and can be linked to each other thus:

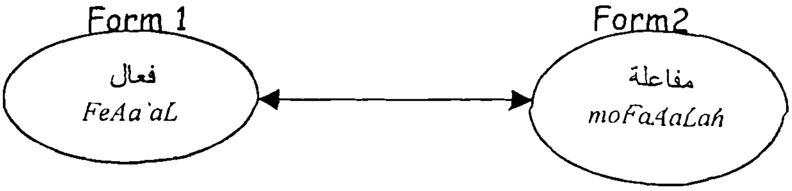


Figure 7.10 semantic link between forms

From the above semantic network, again, hundred of words can be linked to each other Table 7.2 shows some examples of these words.

Table 7.2 example of words based on semantic link

Form 1 (فعال)	Form 2 (مفاعلة)	**English meaning**
قتال	مقاتلة	fighting
نقاش	مناقشة	debate
جدال	مجادلة	argument
جوار	مجاورة	neighbourhood
دفاع	مدافعة	nefence
رهان	مراهنة	wagering, bet

This representation promises to improve the retrieval performance of Arabic information retrieval systems. To test how this method behaves, a pilot study was carried out to compare the performance of this method against the other methods (i.e. word, stem and root). The findings are discussed in Chapter 8.

7.8 Morphological Processor

The morphological analyser uses rules of word formation (i.e. derivation and inflection morphology). This processor is at the heart of the system. It contains several modules, as can be seen from Figure 7.11

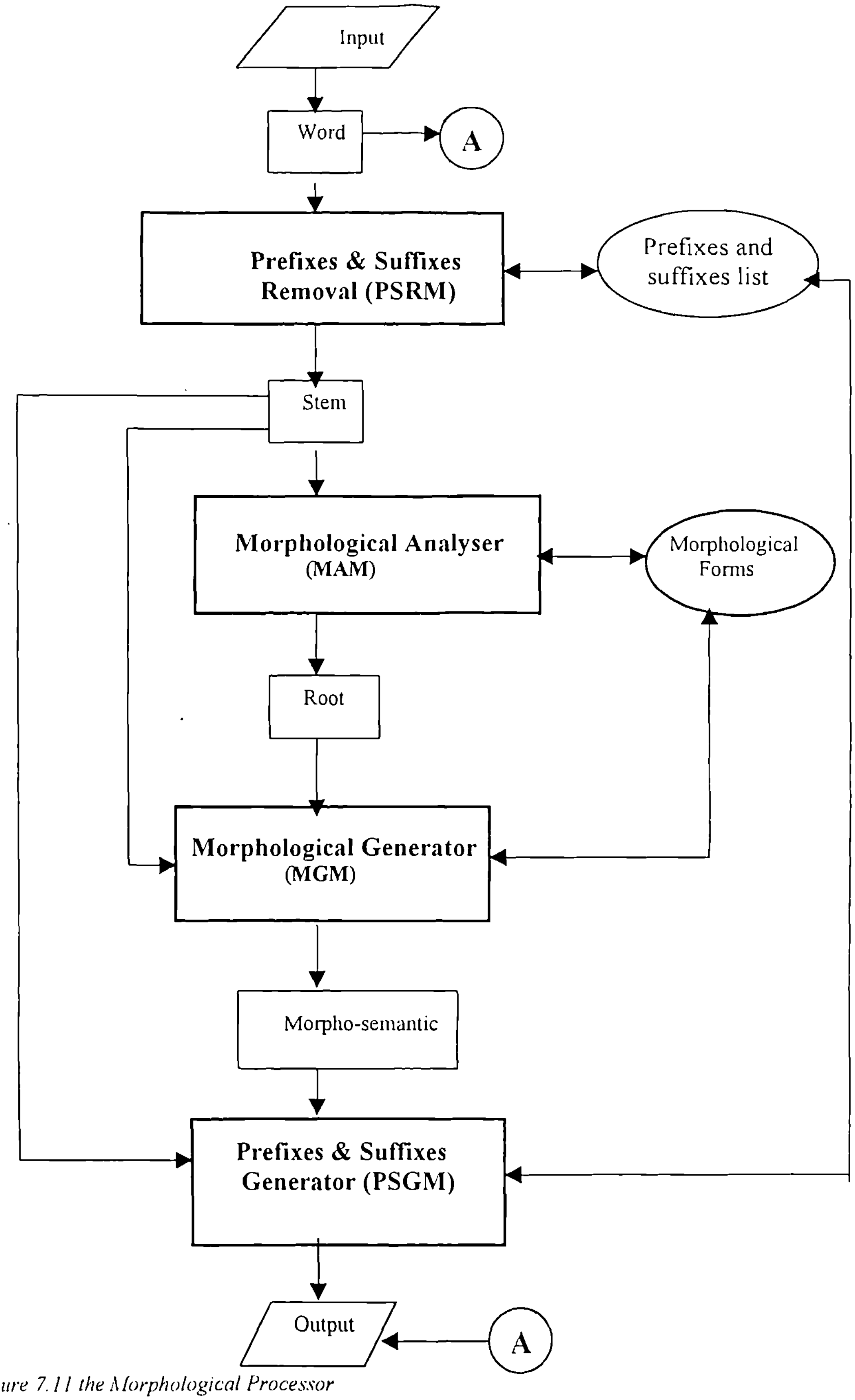

Input
Word
A
Prefixes & Suffixes Removal (PSRM)
Prefixes and suffixes list
Stem
Morphological Analyser (MAM)
Morphological Forms
Root
Morphological Generator (MGM)
Morpho-semantic
Prefixes & Suffixes Generator (PSGM)
Output
A

Figure 7.11 the Morphological Processor

The main aim of developing this processor is to support the three methods of search, namely stem, root and morpho-semantic. It was pointed out here that this processor does not covers all aspects of Arabic morphology, such as phonological changing, verb forms etc. It is not the aim of the study to cover all aspects of Arabic morphology.

7.8.1 Prefixes & Suffixes Removal Module (PSRM)

This module is based on the inflectional theory of Arabic morphology which was mentioned in Chapter 6. The main function of the module is to remove prefixes and suffixes attached to words. To do this, two lists of prefixes and suffixes are developed When a given word is analysed by the PSRM, the following decomposition is produced

Prefixes + Stem + Suffixes
↓
(Root + morphological form)

Then the attached prefixes and suffixes are removed, and the output of this module is a stem only. The stem contains two elements (not separated from each other; in other words, the two elements are integrated): the radical characters (the root or the base of the word), in addition to the morphological forms (or the shape of the word).

Suppose that a user enters the following word: الدراسة (studying). If the word is passed to the prefixes and suffixes removal module (PSRM), the attached prefix ال will be removed. Since there is no suffix in this word the PSRM will do nothing to it. After the prefix ال is removed, the output will be the word stem دراسة. If further analysis is needed, then the output (stem) of this module is passed to the morphological analyser module.

Before leaving this section, it would be helpful to mention the advantages and disadvantages of inflectional theory with respect to its use in information retrieval. In fact there are two main advantages in using inflection theory for information retrieval. Firstly, in most cases, the meaning of a word will not be affected by inflection affixes. Secondly, with inflection affixes, left and right truncation can be used instead of using a more in-depth linguistic analysis. On the other hand, there are two main disadvantages of inflectional theory for information retrieval in Arabic. First, the vast majority of word formation in Arabic is based on derivation and not inflection. Furthermore, inflection affixes often occur with verb forms. Second, with an information retrieval exercise based on inflectional theory many word forms would not be retrieved, due to the change which often occurs within Arabic words (infixes).

7.8.2 Morphological Analyser Module (MAM)

In the previous section the inflectional theory and its implementation was discussed. This section deals with the implementation of derivational theory. Derivational morphology needs deeper analysis than inflection. The main function of the morphological analyser module is to reduce a given stem to its root (base form) from which it was derived. To apply derivational morphology, the MAM was supported by a complete knowledge base of Arabic morphological forms and rules.

The MAM receives the stem (root with morphological form) which was analysed by the PSRM.. In order to reduce the stem to its root, the following steps are followed:

- The MAM needs to consult the morphological forms knowledge base in order to find a match for the stem in the database;
- If a match is found, the MAM removes the augmented characters and vowels from the stem;
- The remaining characters are the root characters (which is, in most cases, three characters only);
- If there is no match, this means that there is no equivalent morphological form for the stem. This means that the stem may not be a correct form or the stem needs to be analysed again.

7.8.3 Morphological Generator Module (MGM)

This module is developed to support the morpho-semantic and root methods. When the given word is reduced to its stem, it will be passed (if there is a need) to the generator module to establish links between related forms. This module is capable of generating several words from a single stem. In order to do this, the morphological forms knowledge base needs to be consulted.

Regarding the root method, this module will generate all forms of the root (stems) which may be relevant or not, while the morpho-semantic method will generate only the relevant ones. When all possible forms of the root are generated through the MGM, the output of this module is passed to the prefixes and suffixes generator module (PSGM).

7.8.4 Prefixes & Suffixes Generator Module (PSGM)

This module adds all prefixes or suffixes that may be attached to the stem. This is the final stage of the processor. The output of this module represents all the forms of the given word entered by the end-user.

7.9 Summary

This chapter discussed the system architecture which was developed in this work. The main aim of developing this prototype is to run and test the study questions about retrieval performance for Arabic. The prototype is based on two Arabic morphological theories: derivational and inflectional. In the chapter the system components were discussed. The system prototype can be used as a search engine for an Arabic text database. Four methods of search are offered by the system. These methods are: word, stem, root and morpho-semantic. The three former methods are available in Arabic commercial retrieval systems, while the last method has been first introduced by this study. The prototype was developed using Prolog. The next chapter will evaluate the system's performance and how the system behaves with respect to each method of search.

Chapter Eight: System evaluation

8.1 Introduction

The previous chapter was a description of the implementation of the prototype (AIRSMA). In order to run the experiments and to evaluate the prototype, a sample of 590 Arabic records was used as a database. The sample was selected to be representative of Arabic texts and was selected from various sources, as was discussed in Chapter 5. The prototype was developed in order to answer the study questions, which were mentioned in Chapter 1 However, this chapter discusses the outcome of the prototype evaluation. It starts with a data representation of the main findings of the evaluation. The main prototype evaluation parameters are covered in the chapter. The study uses four parameters to evaluate the retrieval performance of each method of search (i.e. word, stem, root, and morpho-semantic). The failure analysis for each method is also represented in this chapter.

8.2 Findings

Table 8.1 shows records of the totals retrieved by the four methods (word, stem, morpho-semantic and root). Column 1 shows the number of queries being used in this study. As shown in the table, the total number of queries was 32. The second column of the same table shows the query statement in the Arabic language. (A translation of these queries is listed in Appendix 3.) Column 3 indicates the relevant records available in the database. As was mentioned in Chapter 5 all 590 records were handed to specialist judges to assign relevance. The remaining columns (i.e. 4, 5, 6 and 7) show the number of records retrieved by all four methods of search.

Table 8.1 Total number of relevant records in the database and retrieved records by each method

No.	Query	Relevant	Word	Stem	Morpho	Root
1	المخدرات	3	3	3	3	3
2	النباتات	10	6	18	18	18
3	الطب	4	4	4	7	7
4	الدواجن	5	4	5	5	5
5	الألبان	3	1	2	3	3
6	التربة	7	10	13	19	20
7	المحاصيل	3	1	1	3	6
8	مياه الشرب	4	5	7	7	8
9	الإشعاع	3	3	5	9	9
10	النفايات	7	5	6	6	7
11	تلوث الهواء	8	3	5	9	9
12	حماية البيئة	4	4	8	8	9
13	سوق العمل	11	12	13	13	13
14	التعليم الهندسي	4	3	4	4	4
15	الأستاذ الجامعي	4	3	3	3	3
16	الفهرسة	8	6	9	17	17
17	المجلات الإلكترونية	3	3	5	5	0
18	الإيداع	3	3	4	4	4
19	المكتبات الجامعية	6	3	13	13	19
20	حقوق المؤلف	6	4	5	6	6
21	المكتبة الإلكترونية	3	5	13	13	0
22	حقوق الإنسان	3	4	5	5	5
23	الكتب الممنوعة	5	5	5	5	5
24	مجلة المجتمع	3	3	4	4	4
25	العملات	18	3	19	20	61
26	المصارف	18	9	11	25	25
27	التقسيط	13	7	13	13	13
28	الربا	16	5	10	13	13
29	الديون	12	7	10	19	19
30	الضرائب	7	5	6	7	7
31	العقود	9	9	9	12	14
32	النقود	41	20	20	43	45
Total		**254**	**168**	**258**	**341**	**381**

Consider, as an example, query number 11 about "air pollution". The relevant records for the query in the database number 8 records. As shown on Table 8.1, the word method retrieved only 3 out of the 8 relevant records, while the stem method retrieved 5 records. Both morpho-semantic and root methods retrieved 9 records including the 8 relevant records. As indicated by Table 8.1, the root method generally retrieved more records than the other methods. This followed by the morpho-semantic method comes next. The stem method retrieved more records than the word method did. However, Table 8.1 shows only the retrieved records by each method; it does not tell us how many relevant or irrelevant records have been retrieved by each method.

Since Table 8.1 does not tell us how many relevant and irrelevant records were retrieved by each method the above table needs to be broken down into two sets (Tables): relevant and irrelevant records. To do so, all the records which have been retrieved by each method need to be examined in order to distinguish between the relevant and irrelevant records. After examining all the records, the outcomes of the search methods are represented in Table 8.2 and Table 8.3.

Table 8.2 shows the relevant and irrelevant records retrieved by each method. As far as irrelevant records are concerned, the root method retrieved more irrelevant records than the other methods. The root method retrieved 157 irrelevant records out of 381. In other words, 41% of records being retrieved by the root method were irrelevant. The second method, which brings more irrelevant records (after the root method) was the morpho-semantic method. Out of 341 records retrieved by the morpho-semantic method, it was found that 111 (33%) records were irrelevant. The stem method brings fewer irrelevant records than the root and morpho-semantic methods. The irrelevant records retrieved by the stem method totalled 58 (22%) out of 258 retrieved records. The method that retrieved fewer irrelevant records than the other methods was the word method. It retrieved 31 (18%) irrelevant records out of 168 records.

As far as relevant records are concerned, the morpho-semantic method retrieved more records than the other methods. This performance of the morpho-semantic over the word and the stem methods was expected. However, what was not expected was that the morpho-semantic retrieved more relevant records than the root method did. In theory, root method is expected to retrieve more relevant records than the other methods. However, the output of both methods (morpho-semantic and root) were examined to find out the reasons for such performance. It was found that the root method failed to retrieve any relevant records which contained Arabised words such as إلكترونك, كمبيوتر .

Table 8.2 Total number of relevant and irrelevant records retrieved by each method

	Relevant and retrieved				Irrelevant and retrieved			
No.	Word	Stem	Morpho	Root	Word	Stem	Morpho	Root
1	3	3	3	3	0	0	0	0
2	5	10	10	10	1	8	8	8
3	3	3	4	4	1	1	3	3
4	4	5	5	5	0	0	0	0
5	1	2	3	3	0	0	0	0
6	6	7	7	7	4	6	12	13
7	1	1	3	3	0	0	0	3
8	3	4	4	4	2	3	3	4
9	1	2	3	3	2	3	6	6
10	5	6	6	6	0	0	0	1
11	3	4	8	8	0	1	1	1
12	1	4	4	4	3	4	4	5
13	10	11	11	11	2	2	2	2
14	3	4	4	4	0	0	0	0
15	3	3	3	3	0	0	0	0
16	4	5	8	8	2	4	9	9
17	1	3	3	0	2	2	2	0
18	3	3	3	3	0	1	1	1
19	0	6	6	6	3	7	7	13
20	4	5	6	6	0	0	0	0
21	1	3	3	0	4	10	10	0
22	2	3	3	3	2	2	2	2
23	5	5	5	5	0	0	0	0
24	2	3	3	3	1	1	1	1
25	3	18	18	18	0	1	2	43
26	9	11	15	15	0	0	10	10
27	7	13	13	13	0	0	0	0
28	4	10	13	13	1	0	0	0
29	6	8	10	10	1	2	9	9
30	5	6	7	7	0	0	0	0
31	9	9	9	9	0	0	3	5
32	20	20	27	27	0	0	16	18
Total	**137**	**200**	**230**	**224**	**31**	**58**	**111**	**157**

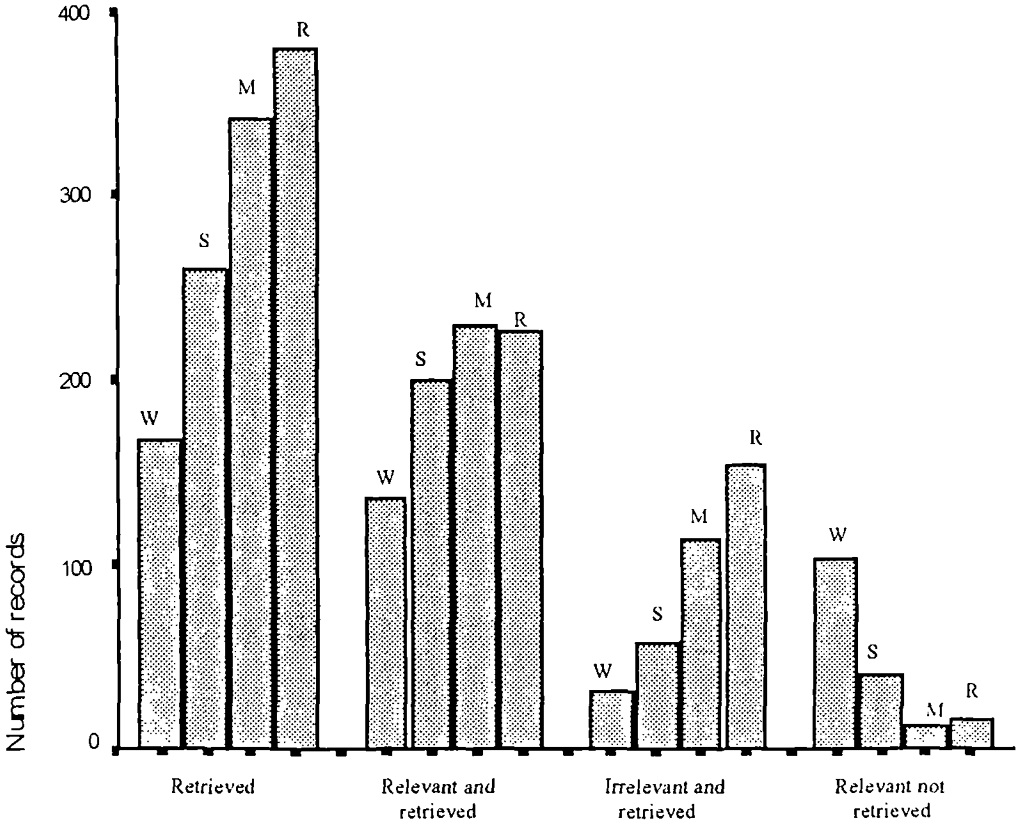
As can be seen from Table 8.2, the root method retrieved 224 (88%) out of 254 relevant records. The morpho-semantic method retrieved 230 (91%) relevant records out of 254, while the stem method retrieved 200 (79%) relevant records. The word method retrieved fewer relevant records than the other three methods. Out of 254 relevant records in the database, the word method retrieved 137 (53%).

Table 8.3 below shows the relevant records which were missed by each method. 117 relevant records, out of 254, were missed by the word method. The stem method missed 54 relevant records out of 254. The morpho-semantic method missed 24 relevant records, while the root method missed 30 relevant records. It was found that each relevant record retrieved by the word, stem and morpho-semantic methods was retrieved by the root method except those words which were converted from other languages, such as English.

Table 8.3 Total number of relevant records missed by each method

No.	Word	Stem	Morpho	Root
1	0	0	0	0
2	5	0	0	0
3	1	1	0	0
4	1	0	0	0
5	2	1	0	0
6	1	0	0	0
7	2	2	0	0
8	1	0	0	0
9	2	1	0	0
10	2	1	0	1
11	5	4	0	0
12	3	0	0	0
13	1	0	1	0
14	1	0	0	0
15	1	1	1	1
16	4	3	0	0
17	2	0	0	3
18	0	0	0	0
19	6	0	0	0
20	2	1	0	0
21	2	0	0	3
22	1	0	0	0
23	0	0	0	0
24	1	0	0	0
25	15	0	0	0
26	9	7	3	3
27	6	0	0	0
28	12	6	3	3
29	6	4	2	2
30	2	1	0	0
31	0	0	0	0
32	21	21	14	14
Total	**117**	**54**	**24**	**30**

Figure 8.1 gives a summary of the performance of each method. It represents all data in Tables 8.1, 8.2 and 8.3. The behaviour of each method is clear from this figure whether concerning records being retrieved or whether retrieval were relevant or irrelevant.



W= word method
S = stem method
M = morpho-semantic method
R = root method

Figure 8.1 retrieval performance of each method

However, it is not possible to draw a more accurate conclusion about the retrieval performance of each method from the above Tables. To do so, retrieval performance parameters are needed (such as recall and precision), the two parameters which are commonly used in retrieval system evaluation. In this study as mentioned in Chapter 5, these two parameters are used in addition to two other parameters (i.e. noise or false drop and the omission factor).

8.3 Retrieval performance parameters

The outcome of each method was tested using four parameters. Two of them are well known in the retrieval evaluation environment. In this section, a brief description of each parameter is given.

8.3.1 Recall and precision measures

Recall (*R*) and precision (*P*) are the most important and well-known parameters in the field of information retrieval evaluation. The recall parameter measures how well the system retrieves all the relevant records. In other words, recall is the ratio of relevant records retrieved for a given query over the number of relevant records for that query in the database. A precision parameter measures how well the system retrieves only the relevant records. In other words, precision is the ratio of the number of relevant records retrieved over the total number of records retrieved. They are computed as follows:

$$R \qquad \text{RETREL} / (\text{RETREL} + \text{NRETREL})$$

$$P \qquad \text{RETREL} / (\text{RETREL} + \text{RETNREL}),$$

Where RETNREL is defined as the number of records retrieved but not relevant, RETREL is the number of records retrieved and relevant, and NRETREL is the number of records relevant but not retrieved. When the recall and precision parameters are used to test the retrieval performance of each method, the output of the two parameters is shown in Table 8.4.

Table 8.4 mean recall and precision for each method

Method	Mean Recall	Mean Precision
Word	54%	82%
Stem	79%	78%
Morpho-semantic	91%	67%
Root	88%	59%

Table 8.4 shows the mean recall and precision for each method. As can be seen from the table, the morpho-semantic method achieved the highest recall against the other methods at a level of 91%. The highest precision was achieved by the word method at 82%. However, the higher the recall which is achieved, the more of the level precision decreases. As Lancaster points out (Lancaster, 1978), recall and precision tend to be related inversely. This means that to achieve better recall, precision will tend to go down. Conversely, when a great precision is needed, recall will tend to deteriorate.

As can be seen from Table 8.4, the retrieval performance of recall for the morpho-semantic method is superior to the other methods. The superiority of the morpho-semantic method over the word and stem methods was expected. What was not expected was the superiority of the morpho-semantic method over the root method. As shown in Table 8 4, the recall of the morpho-semantic method was at a level of 91%, while the recall for the root method was at 88%. In theory, it was expected that the root method would retrieve more relevant records than the other methods (including the morpho-semantic method). This expectation was based on the fact that the root method is capable of retrieving most, if not all, of the morphological variations of a given word. However, after examining and comparing the records being retrieved by each method, it was found that the reason for such a performance for both methods was related to the non Arabic word being used in a query or texts such as المجلات الالكترونية (electronic journals). In other words, the morpho-semantic method was successful in retrieving them, while the root method was not.

Concerning the precision performance for each method, Table 8.4 shows that the highest precision was achieved by the word method at a level of 82%, while the lowest level was achieved by the root method at 59%. Again, each parameter has an effect on each of the others for the word, morpho-semantic and root methods. The stem method achieved similar levels for both parameters: 79% for recall and 78% for precision. As can be seen from Table 8.4, the performance of the morpho-semantic method for precision was superior to that of the root method. The table shows that the precision for the morpho-semantic method was at a level of 67%, while the root method was at a level of 59%.

In the previous sections, a discussion was represented on the retrieval performance of each method in terms of recall and precision. However, the causes or reasons for this performance was not discussed; for example, why the word method has the lowest recall among the other methods and vice versa, why the morpho-semantic method has the highest precision. In order to answer these questions and others, a failure analysis was used and is discussed later on in this chapter.

8.3.2 Noise or false drop measure and omission factor

The noise or false drop (*N*) is used as the ratio of irrelevant records retrieved to the total retrieved. It is also known as the complement of precision. The second parameter is the omission factor (*O*), which is defined as the portion of relevant records not retrieved. This parameter is also called the complement of recall or the conditional probability of a miss. The two parameters are computed as:

$$N \quad \text{RETNREL} / (\text{RETREL} + \text{RETNREL})$$
$$O \quad \text{NRETREL} / (\text{RETREL} + \text{NRETREL})$$

The above measures were used for each method; the results are shown below in Table 8.5.

Table 8.5 Noise or false drop and Omission factor parameters for each method

Method	Noise or false drop	Omission factor
Word	18%	46%
Stem	22%	21%
Morpho-semantic	33%	09%
Root	41%	12%

As shown above, the output of the noise parameter (or retrieving more irrelevant records than other methods) for the root method shows the highest level (41%) when it is compared with other methods. The lowest level of noise or false drop was achieved by the word method (i.e. retrieving fewer irrelevant records than other methods) at a level of 18% However, as far as the omission factor is concerned, the highest level was achieved by the word method (46%) (missing more relevant records than other methods); and the lowest level of omission factor was achieved by the morpho-semantic method (i.e. missing fewer relevant records than other methods) at a level of 09%.

It seems, as shown in Table 8.5, that when the noise or false drop has achieved a low level of noise, this achievement has an effect on the omission factor. When the output of the two parameters is compared one to the other, a significant difference between the output of each parameter can be noted. For example, when the noise or false drop parameter is considered for the word method, it was found that the word method behaves well (i.e. a small number of irrelevant records were retrieved). In other words, the ratio of noise or false drop was 18%. This superiority concerning the noise parameter of the word method has an effect upon the omission factor of the same method. In brief, the level of the omission factor goes up to a level of %46. This is also true for the morpho-semantic and root methods. However, the stem method achieves a similar ratio for both parameters: 22% for noise or false drop and 21% for the omission factor. This means there is no significant effect of noise or false drop over the omission factor where the stem method is concerned.

Before the discussion of the failure analysis, the 32 queries and their output will be divided into five groups according to subject coverage. This type of division of the outcome may tell us of how each method behaves under each subject matter. However, the study does not aim to discuss the effect of the subject coverage on retrieval performance rather than giving an overview of its impact. Table 8.6 shows recall and precision levels of each method under different subject coverage.

Table 8.6 The retrieval performance of each method according to subject coverage

		Recall				**Precision**			
Query Number	Subject coverage	Word	Stem	Morpho	Root	Word	Stem	Morpho	Root
1-12	Group 1	60%	83%	98%	98%	73%	66%	62%	58%
13-15	Group 2	84%	95%	95%	95%	88%	90%	90%	90%
16-21	Group 3	45%	86%	100%	79%	54%	51%	60%	50%
22-24	Group 4	81%	100%	100%	100%	75%	79%	79%	79%
25-32	Group 5	47%	71%	84%	84%	37%	37%	33%	29%
Average		**63%**	**87%**	**95%**	**91%**	**65%**	**64%**	**65%**	**61%**

Group 1: Science and Environment
Group 2: Education
Group 3: Library and Information Science
Group 4: General subject
Group 5: Economics

8.4 Failure analysis

Having derived the retrieval performance of each method in the previous tables, the next step involved a detailed intellectual analysis of reasons why recall and precision failures occurred. This type of analysis is used in order to answer the following question: why did recall and precision failures occur? What are the reasons for the failure? In fact there are a number of reasons for each failure, Lancaster (1969) and (Jones, 1999) listed most of them. For example, the recall failure may have occurred because search formulation was too exhaustive or too specific; it may also have occurred because of the indexing policy, morphological variations and so on. Precision failure may be due to the fact that terms are near but not related, mentioned only in passing or because of common non-content words. The next section discusses the failure analysis of recall and precision for each method.

8.4.1 Recall failure

The recall failure analysis can explain why a given method failed to retrieve relevant records even though they were in the database. As discussed above, the word method failed to retrieve 117 relevant records out of 254. The stem method also failed to retrieve 54, the morpho-semantic method failed to retrieve 24, and the root method 30 relevant records. After examining all records for each method, the cause of failure can be divided into the following categories:

Table 8.7 Type of recall failure for each method

Type of failure	Word	Stem	Morpho	Root
Morphological variations	93	30	-	-
Synonyms	24	24	24	24
Non Arabic word	-	-	-	6
Total failures	**117**	**54**	**24**	**30**

8.4.1.1 Morphological variations

The main recall failure for word and stem methods was due to the fact that the query used for searching the database appears in some records in a variety of morphological variations, as shown in Table 8.7. Of the 117 recall failures for the word method, it was found that 93 (79%) of these corresponded to morphological variations and of the 54 recall failures for the stem, it was found that 30 (56%) corresponded to morphological variations This type of failure does not occur for the morpho-semantic and the root methods. This is due to the ability of both methods to search for the morphological variations of the given keyword that may occur in the database.

8.4.1.2 Synonyms

The four methods failed to retrieve 24 relevant records. This failure occurred because of synonyms. Of the 117 recall failures for the word method, it was found that 24 (21%) of these corresponded to synonyms. The stem method failed to retrieve 24 of the relevant records (44%) out of 54. All the recall failures for the morpho-semantic method were due to synonyms, while the root method failed to retrieve 24 (80%) relevant records out of 30. In fact, this type of failure is related to semantic analysis rather than morphological analysis, and this failure is beyond the scope of this study.

Since the synonyms beyond the scope of the study, it may be preferable to investigate how each method performance without including synonyms in the analysis. To do this let us, first, recall the data in Tables 8.4 and 8.5. From Table 8.4, the recall level of the word method was 54%, the stem method was 79%, the morpho-semantic was 91% and the root method was 88%. Furthermore, from Table 8.5 the omission factor level for the word method was 46%, the stem was 21%, the morpho-semantic method was 09% and the root method was 12%. However, if the synonyms are removed form the analysis, the retrieval performance of each method, as far as recall is concerned, will be increased, while the omission factor level will be decreased as shown below in Table 8.8.

Table 8.8 Retrieval performance for each method without synonyms

Method	Mean Recall	Omission factor
Word	60%	40%
Stem	87%	13%
Morpho-semantic	100%	0
Root	.97%	03%

If the data in the above table is compared with data in Tables 8.4 and 8.5, one can notes the different performance for each method.

8.4.1.3 Non Arabic words or Arabised words

This failure occurs only when the root method is used. As mentioned above, out of 30 relevant records missed by the root method, it was found that 24 records were synonyms. After examining the remaining records (6), it was found that the cause of failure related to a non-Arabic word, such as المجلات الالكترونية (electronic journals) being used by the query and the records themselves. This means that the root method is unable to deal with any records or queries containing non-Arabic words. However, previous experiments, such as those of Hmeidi (1995) use non-Arabic words (Arabised word) as roots. Hmeidi did not build his system based on morphological analysis. In other words, extracting the root of a word was carried out manually, while in this study, reducing a given word to its root was, carried out automatically.

Why high recall?

From the above tables it has been seen that the recall level of morpho-semantic and root methods have achieved a higher level than the word and stem methods. It would be asked why the two methods have a higher recall level than the other methods. The answer is that the two methods (i.e. the morpho-semantic and the root method) are capable of retrieving morphological variations of a given word, while the other two methods are not. This capability of both methods is based on the morphological analysis that supports the two methods.

Why high precision?

The opposite question is to ask why the word and stem methods have a higher precision than the other methods. This is because the morpho-semantic and root methods used morphological analysis which is able to retrieve more forms of a given word. This, in turn, may bring with it more irrelevant words, while the word and stem methods is deals with the word itself (in the case of the word method), and remove prefixes or suffixes of a given word (in the case of the stem method).

8.4.2 Precision failure

This measure is used to find out about the performance of each method with respect to retrieving only the relevant records in the database, and rejecting the irrelevant records at the same time. It was found that this type of failure is very rare when the word, stem and the morpho-semantic methods are used. On the other hand, this failure increases when the root method is used, as shown in Table 8.9. Precision failures can be subdivided into the following types:

Table 8.9 Type of precision failure for each method

Type of failure	Word	Stem	Morpho	Root
Morphological variations	-	-	-	46
Homographs	-	1	28	28
Semantic context	-	-	11	11
Relevance judgment	31	57	72	72
Total of failure	**31**	**58**	**111**	**157**

8.4.2.1 Morphological variations

As can be seen from Tables 8.7 and 8.9, the morphological variations occur for both recall and precision. The failure, here precision failure, is the opposite of the recall failure of morphological variations. To differentiate between the two failures, the following example may be considered. Suppose a user wants to search for the keyword "العُمْلَة", which means currency. If the root method is used, the query is reduced to the root عمل . From this root several forms can be generated, such as عمال (labours), عملية (operation), عملة (currency), عوامل (factors) and so on.

As shown in Table 8.8 the precision failure for the root method was 46 (29%) records out of 157 which corresponded to morphological variations. This means that the root method generates and retrieves more irrelevant records than the other methods.

8.4.2.2 Homographs

Homographs are forms which differ phonetically but are spelled in the same way This type of failure occurred only when the morpho-semantic and the root methods were used. It was found that 28 records out of 111 of the precision failures for the morpho-semantic method were due to homographs. The root failure was 28 records out of 157. An example of a homograph which was found by the study is the word (دين). Query number 29 was about the loans (ديـــون). When the root method is used, the word (ديون) was reduced to its root (دين). The root has two meanings as follows:

1- (دَيْن) singular of loans (ديون), and

2- (دِين) which means religion.

Both meanings were retrieved by each method (i.e. morpho-semantic and root), though there is no relationship between loans and religion.

8.4.2.3 Semantic context

This failure also occurred when the morpho-semantic and root methods were used. This type of failure depends on the context of the word (i.e. the meaning of the word cannot be understood without the context) that may appear in the text. For example, in query number 32 ("النقـــود" means money) the morpho-semantic method retrieved 43 records, while the relevant records in the database for query 32 numbered 41.

After examining the records being retrieved, it was found that out of the 43 retrieved records, the relevant records totalled 27. The remaining records (16) were irrelevant. Further examination of the 16 records found that 11 records were about (النقد),

which means criticism. Also the same word can be used as the verbal noun "نَقَد" of the word "money".

8.4.2.4 relevance judgement

The failure here related to human decision rather than system failure. In other words, the method may be succeeding to make a good match between a query and the relevant records. To make this clear is necessary to consider query number 12 about "environment protection". The word method retrieved 8 records, while the number of relevant records in the database totally only 4. When the 8 records were examined, it was found that in addition to the 4 relevant records, the same terms used in the search appear in the other 4 records, but judges assigned the other 4 records irrelevant. Further examination of these records found that the term "environment protection" was been mentioned in these records only in passing. As shown in Table 8.9, all the precision failures for the word and stem methods were related to relevance judgements. In other words, the two methods succeeded in achieving an exact match between the query and the terms which appeared in the records but the judges assigned these records as irrelevant. The Morpho-semantic and root methods have the same number of failures related to relevance judgements.

8.5 Summary

The main aim of the empirical study was to compare the retrieval performance of the morpho-semantic method against the three methods of search used in information retrieval for Arabic: namely, word, stem, and root. An information retrieval prototype was developed using Prolog. An important feature of this work is the introduction of the morpho-semantic method (a novel approach) to be used in information retrieval for the first time. The results give a clear indication that the morpho-semantic method has great potential for improving the retrieval performance of the word and stem methods, especially for recall. The same method has also improved the retrieval performance of the root method, especially in terms of precision. This is by no means a claim that the morpho-semantic method is better or worse than the other methods as far as the sample data and the limitations of this study are concerned. I believe that the morpho-semantic method is a promising method for use in Arabic information retrieval systems.

Chapter Nine: Summary and conclusion

9.1 Introduction

The purpose of this chapter is to summarise the achievement which has been made by the study. This includes a review of the research design. It is then followed by the main findings. Based on these findings, a number of recommendations are represented in this chapter. Finally, the chapter ends with some future work which is needed to be carried out as a continuation of this study.

9.2 Study summary

This section aims to draw the attention of the reader to what the study is about, in other words, what the study is trying to achieve. In brief, information retrieval systems in Arabic are few. Some of these systems were Arabised from Latin languages, mostly English, while other systems are fully Arabic systems. However, developing an information retrieval system in the Arabic is a new phenomenon. To do so, many characteristics of Arabic language must be taken into account, especially morphological analysis. As it is known in Arabic, there may be some, or many, different forms of a given word, these forms resulting from the addition of different prefixes, suffixes or infixes to a common word stem according to the dictates of grammar.

In Arabic information retrieval systems, three search methods are used namely, word, stem, and root. The word method is based only on term matching, while the other two methods are based on morphological analysis, (they have different levels of morphological analysis). However, each method has its limitations. For example, the word and stem methods may miss relevant records (because of morphological variations), while, on the other hand, the root method may retrieve irrelevant records. This due to the fact that the root method is capable of reducing a given word to its root, and it will then generate all possible morphological variations of that word.

Having identified some limitations of the current search methods in Arabic, the present study has introduced a novel approach based on what is called the morpho-semantic method. The development of this method is based on the hope that it may improve the effectiveness of the word and stem methods in terms of retrieving more relevant records and, at the same time, it is hoped that the same method will improve the root method in terms of rejecting irrelevant records that may be retrieved by this method. This approach has used a representative sample of Arabic morphological forms (nouns and adjectives) which form the basis of the majority of the Arabic words.

To address the problems identified above, a prototype information retrieval system was developed using Prolog. The prototype was based on four search methods: namely, word, stem, root, and the morpho-semantic method. The first methods (i.e. word, stem and root) have been used in some commercial software, as well as ina number of academic studies, while the morpho-semantic method has yet not been used (as far as the writer of this thesis knows). Moreover, although the stem method is used in some commercial software, it has still not been exploited to its full potential.

The morpho-semantic method is based on linking related morphological forms to each other. To implement this, two types of knowledge representation techniques were used: the semantic networks and the rule base. Within the semantic networks, a number of links have been made between related morphological forms such as broken plural and singular, or verbal nouns and agent nouns. After all the work was completed, a number of experiments were carried out to evaluate the retrieval performance of each method. The next section highlights the main outcomes of the evaluation.

9.3 Summary of findings

In Chapter 1, it was mentioned that, in addition to the main aims of the study, there were a number of questions which were asked. This section aims to answer those questions. These answers are based on the results and the outcome of the evaluation.

1- The first question was, does the morpho-semantic method improve the retrieval performance of the word and stem methods in terms of recall? To answer this question a comparative study of retrieval performance for each method was carried out. The results of the comparison show that the retrieval performance of the recall parameter for the morpho-semantic method was at a level of 91%, while the word method was at 54% and the stem method was at 79%. The results show the morpho-semantic method did improve the retrieval performance of the word and stem methods. In other words, more relevant records would be retrieved by the morpho-semantic method when compared with the word and the stem methods. On the other hand, the same method did not improve the retrieval performance of the two methods in terms of precision (i.e. the morpho-semantic method retrieved more irrelevant records than the two other methods). However, this type of performance was not considered in this study.

2- The second question was, does the morpho-semantic method improve the retrieval performance of the root method in terms of precision? Again, from the evaluation results it was found that the precision of the retrieval performance for the morpho-semantic method was at the level 67%, while the root method was at the level 59%. In other words, the morpho-semantic method succeeds in rejecting more irrelevant records than the root method. This means that the morpho-semantic method did improve the retrieval performance of the root method in terms of precision.

3- The third question was: do the morphological variations have an effect on retrieval performance in Arabic? The results show that the morphological variations may have a great affect on retrieval performance when the word method is used. It was found that the word method retrieved only 137 relevant records, while the relevant records in the database number 254. In other words, the word method missed 117 records. The failure of the word was due to morphological variations.

The above show one side of the effects of morphological variations on retrieval performance. The other side is related to retrieving irrelevant records. It was found that morphological variations have an effect on the retrieval performance of the root method. The failure rate related to morphological variations for the root method was 46 irrelevant records, selected out of 157.

4- It is necessary to ask why some methods retrieved more relevant records than the others, or why some methods retrieved more irrelevant records than the other methods did. From the results, it is indicated that with reasonable morphological analysis (i.e. removing prefixes or suffixes from a given word) more relevant records would be retrieved. However, reducing a given word to its root, then generating from that root all possible morphological variations, results in the retrieval of more irrelevant records.

9.4 Limitations

Any conclusions drawn from the findings of the study should be interpreted within the limitation of the study design. As mentioned above, the study attempted to apply a novel approach (morpho-semantic) to be used in Arabic information retrieval. This method is compared against three methods of search: namely, word, stem and root. The morphological analysis techniques used in this study are restricted to Arabic text.

One of the limitations of the study is the test collection being used. The study used 590 Arabic records, in addition to 32 queries to search these records. There is no need to give further details about the collection since this is covered in Chapter 5. However, an attempt was made to cover a number of subject areas, and the results cannot be generalised to other subject areas. Therefore, the conclusions should be made only in the context of the test collection of the study.

Another limitation of the current study is that the morphological analyser used in this study is not completed. It is expensive to develop a full morphological analyser and generator in Arabic. This is because of the complexity of the Arabic language's structure. However, the morphological analyser and generator being used in this study has succeeded in carrying out the tasks.

A final limitation was that, the study used only one approach of linguistic analysis (i.e. morphological analysis). Other linguistic approaches, such as syntactic, semantic, and pragmatic analyses are beyond the scope of the study.

9.5 Recommendations

The effectiveness of the morphological analysis on the retrieval performance was noted by the current study. This led the researcher of the study to address some recommendations related to this issue:

1- An Arabic test collection of texts with queries is needed. The test collection being used by the study can be used as base for developing an adequate test collection for information retrieval experiments in Arabic;

2- More Arabic retrieval systems are needed in order to use them as experimental systems, such as SMART in English or Mocro-AIRS in Arabic. Unfortunately, no ready-made experimental Arabic retrieval system is available for Arab researchers.

3- The researcher hope for the following groups: Computer, Libraries and Information Departments to direct some of their work towards information retrieval in Arabic;

4- Based on the literature review, it was found that there is a weakness of cooperation between three groups of scientist (i.e. computer scientists, librarians, and linguists). As a result of that, it might helpful, for improving the work on this area, if the three groups could work together to enhance information retrieval performance in Arabic;

5- The research desires the non profit organisation such as KFNL and KACTS or other organisations to support researches on IR through offering some grants;

6- Numbers of artificial techniques such as inelegant information agents and natural language processing may be used to enhance the retrieval performance of Arabic systems;

7- Because of the increase of the Arabic sites on the Internet, the advance search engines are essential to support the end user with their needs.

9.6 Future work

The study results of the morpho-semantic method have motivated the researcher to carry out more work in this area. This involves the extension of the work on semantic linking of morphological forms. The prototype, which was used in this study could be improved via in-depth analysis of Arabic morphology. The researcher is also planning to investigate the Multi Agent System techniques to support morphological analysis for Arabic information retrieval systems.

Another option of future research is that of conducting large sample texts with real search requests to validate the results of the present study. Using this sample as a test collection may enable the researcher to prove the value of using the morpho-semantic method for information retrieval in Arabic.

Finally, I hope to undertake the following works either by myself or as part of a team working on the following:

1- Morphological analysis theories need to be investigated and developed in order to apply them to information retrieval in Arabic.

2- The broken plural has a great effect on information retrieval in Arabic. This type of number and its effect on information retrieval has not been studied yet.

3- A study is needed to compare the retrieval performance of word-based prefixes only with word-based suffixes only.

4- The semantic relationship between the morphological forms needs to be investigated.

5- More studies are needed to improve retrieval performance in Arabic through morphological analysis.

6- A comprehensive literature review and bibliography for Arabic morphological analysis is needed.

References

AbdAla'al, AbdAlmeeniam (1977). *Sound and broken plurals in Arabic language*. Cairo: Maktabat Al Khanji.

Abu Salem, Hani (1992). *A microcomputer based Arabic bibliographic information retrieval system with relation thesaurus* (Arabic-IRS), Ph.D. Thesis. Chicago: Illinois Institute of Technology.

Al Asa'ad (1993). “الوجيز في التعريف بـــالصرف وتاريخـــه” *Alwajez in morphology and its history*. Riyadh: Dar Al Meearaj.

Al Astee, Abdullah Mohammed (1992). “التعريف في علم التصريف “ *Introducing morphology*. Tarabuls: Kuliat aldawah alislamiah.

Al Atram, Mohammad (1989). “كفاءة اللغة الطبيعية في تكشيف واسترجاع الوثـــائق العربيـــة “ *The effectiveness of the natural language for indexing and retrieving Arabic documents*. Riyadh: KACST.

Al Bakhit, Bakhit Suliman (1993). “البحث في العنوان في قواعد البينات العربية “ Searching titles in Arabic database. In: *Proceedings of Symposium on Using Arabic Language in Information Technology*, pp.569-580.

Al Dosary, Fahd M. and Abdulrahaman H. Ekrish (1991). The state of automation in selected libraries and information centres in Saudi Arabia. *Libri*, 41 (2), pp.109-120.

Al Fedaghi, Sabah S. and Fawaz S. Al Anzi (1989). A new algorithm to generate Arabic root pattern forms. In: *The 11th National Computer Conference*. pp 391-400.

Al Hamlawi, Ahmed (1991). “شذ العرف في فـــن الصـــرف “ *Shatha aloarf in morphology art*. Cairo: Maktabat Alaadab.

Al Jabri, Saad and Chris Mellish (1994). Generating Arabic words from semantic descriptions. In: *Proceedings of the 4th International Conference and Exhibition on Multi-lingual Computing*, 9-6-1.

Al Kharashi, Ibrahim (1991). *A microcomputer-based Arabic information retrieval system comparing words, stems, and roots as index terms*, Ph.D. Thesis. Chicago: Illinois Institute of Technology.

Al kharashi, Ibrahim and Martha W. Evans (1993). Comparing words, stems, and roots as index terms for an Arabic information retrieval system. In: *Arabic language and information technology*.

Al Khuli, Ali (1982). "التراكيب الشائعة في اللغة العربية " *Common structure in Arabic language.* Riyadh: Dar Alolum.

Al Naim, Faisal (1989). *Text analysis and automatic indexing for Arabic based automated information retrieval system.* MSc Thesis, Chico: California State University

Al Sawaydan, Nasser (1993). " الاسترجاع الموضوعي بواسطة كلمات العنوان " Subject retrieval via title keywords .In: *Proceedings of Symposium on Using Arabic Language in Information Technology*, pp.533-568.

Al Tayyar, Musaid and Kamal Bechkoum (1998). The effectiveness of the morphological analysis for text retrieval in Arabic. In: *Proceedings of the 6th International Conference and Exhibition on Multi-lingual Computing,* 2.4.1.

Al Swaynia, Ali (1994). " استرجاع المعلومات في اللغة العربية " *Information retrieval in Arabic language*. Riyadh: King Fahd National Library.

Al-Uthman, AbdAlziz Ahmed (1989). *A morphological analyser for Arabic.* MSc Thesis, Dahran: King Fahd Uinversity of Petroleum and Minerals.

Ali, Nabil (1992). Parsing and automatic diacritization of written Arabic: a breakthrough. In: *Proceedings of the 13th National Computer Conference*, pp. 794-812.

Ali, Nabil, (1988). " اللغــة العربيــة والحاســوب " *Arabic language and computer*. Kuwait. Ta`reep.

Aman, Mohammed (1984). Use of Arabic in computerised information interchange. *Journal of the American Society for Information Science,* 35 (4), pp. 204-210.

Anwar, Mohamed Sami (1989). Computer-based lexicography: how much grammar can be included?. In: *Proceedings of the Seminar on Bilingual Computing in Arabic and English.*

Arampatzis, A. T., T. Tsoris, C. H. A. Koster and Th. P. Van Der Weide (1998). Phase-based information retrieval. *Information Processing & Management* 34 (6), pp. 693-707.

Artandi, Susan (1976). Machine indexing linguistic and semiotic implications. *Journal of the American Society for Information Science,* 27 (4), pp. 235-239.

Ashoor, Mohammad Saleh (1989). Arabisation of automated library systems in the Arab world: need for compatibility and standardisation. *Libri,* 39 (4), pp.294-302.

Bessley, K.; Buckwalter and S. Newton (1989). Two-level finite-state analysis of Arabic morphology. In: *Proceedings of the First Conference on Bilingual Computing in Arabic and English.*

Bessley, K (1998). Arabic morphological analysis on the Internet. In: *Proceedings of the 6th International Conference and Exhibition on Multi-lingual Computing, ICEMCO-98.*

Booth, L. M.; Khalid M. Niaz and H. M. Al Swaidan (1986). Arabisation of an automated library system. In: *The Ninth National Computer Conference and Exhibition*, pp.10-32.

Boyce, Bert R.; Charles T. Meadow and Donald H. Kraft (1994). *Measurement in information science.* San Diego: Academic Press.

Buchanan, Brian (1976*). A glossary of indexing term*s. London: Clive Bingley.

Cleverdon, C. W. (1966). *Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems*, 2 vols., Cranfield: College of Aeronautics.

Clocksin, W. F. and C. S. Mellish (1994). *Programming in Prolog.* Berlin: Springer-Verlag

Computer Guide. *IRSAD* (1993). 5 (45), pp.1,39 and 45.

Crystal, David (1991a). *A dictionary of linguistics and phonetics.*

Crystal, David. (1991b). *An encyclopaedic dictionary of language and languages.*

Daraz, Tantawi Mohammed (1986). " ظاهرة الاشتقاق في اللغة العربية " *Derivational phenomenon in Arabic language* . Cairo: Matbaat Abideen.

DeSalvo, William (1992). *Measurement of full text retrieval effectiveness and relevance judgements a cross-varying levels of expertise.* MSc Dissertation. University of North Carolina, School of Information and Library Science.

Doszkocs, Tames (1986). Natural language processing in information retrieval. *Journal of the American Society for Information Science.* 37 (4), pp. 191-196.

El Sadany, T. A. and M. A. Hashish (1989). An Arabic morphological system. *IBM System Journal*, 28 (4), pp. 600-612.

Feddag, Allel (1992). Arabic morpho-syntax and semantic parsing. In: *Proceedings of the 13th National Computer Conference*, pp.717-749.

Feinberg, Hilda (1973*). Title derivative indexing techniques: a comparative study.* Metuchen: The Scarecrow Pree.

Fidel, Raya and Dagobert Soergel (1983). Factors affecting online bibliographic retrieval: a conceptual framework for research. *Journal of the American Society for Information Science.* 34 (3), pp. 163-180.

Finlay, Janet and Alan Dix (1996). *An introduction to Artificial intelligence*. London: UCL Press Limited.

Frakes, William B and Ricardo Baeza-Yates (1992). *Information retrieval: data structures & algorithms*. New Jersey: Prentice Hall.

Fromkin, Victoria and Robert Rodman (1998). *An introduction to language*. New York: Harcourt Brace College Publishers.

Gordon, Michael and Praveen Pathak (1999). Finding information on the World Wide Web. the retrieval effectiveness of search engines. *Information Processing and Management* 35 (1999), pp. 141-180.

Harman, Donna (1991). How effective is suffixing? *Journal of the American Society for Information Science*, 42 (1), pp. 7-15.

Harter, Stephen P. (1986). *Online information retrieval: Concepts, principles, and techniques*. Orlando: Academic Press INC.

Haverkmp, Donaa S. and Susan Gauch (1998). Intelligent information agents: review and challenges for distributed information sources. *Journal of the American Society for Information Science*, 49 (4) pp. 304-311.

Hegazi, N. H. and A. A. El Sharkawi (1985). A computerised lexical analyser for natural Arabic text. In: *Computer processing of the Arabic language*.

Hegazi, N. H. and A. A. Elsharkawi (1986). Natural Arabic language processing. In: *The Ninth National Computer Conference and Exhibition*.

Hilal, Yahiah (1985). Morphological analysis of Arabic speech. In: *Computer processing of the Arabic language*.

Hmeidi, Ismael Ibrahim (1995). *Design and implementation of automatic word and phrase indexing for information retrieval with Arabic documents*, Ph.D. Thesis. Chicago Illinois Institute of Technology.

Hmeidi, Ismael, Ghassan Kanaan and Martha Evans (1997). Design and implementation of automatic indexing for information retrieval with Arabic documents. *Journal of the American Society for Information Science* 48 (10), pp. 867-881.

Honderich, Ted ed. (1995). The Oxford companion to philosophy. Oxford: Oxford University Press.

Hull, David (1996). Stemming algorithms: a case study for detailed evaluation. *Journal of the American Society for Information Science*, 47 (1), pp. 70-84.

Jones, Kendra (1999). Linguistic searching versus relevance ranking: DR-LINK and TARGET. *Online & CD-ROM Review*, 23 (2), pp. 67-80.

Kassem, Neezar (1988). " خصوصيات الأسماء والصفات العربية وأثرها في خزن المعلومات واسترجاعها " Peculiarities of Arabic nouns and adjectives and their effect in information storage and retrieval, *A`adab Almostanseeriah Journal*, 16, pp.705-737.

Katamba, Francis (1993). *Morphology*. London: Macmillan Press Ltd.

Klavans, J. L. and E. Tzoukermann (1992). Morphology. In: Stuart C. Shapiro, ed. *Encyclopaedia of artificial intelligence*. New York: John Wiley & sons.

Lancaster, F. Wilfrid (1968). *Information retrieval systems: Characteristics, testing, and evaluation*. New York: John Wiley & Sons, Inc.

Lancaster, F. W. (1978). Precision and recall. In: Allen Kent eds., *Encyclopedia of Library and Information Science*, vol. 23.

Lancaster, F. Wilfrid and Amy J. Warner (1993). *Information retrieval today*, Arlington Information Resources Press.

Lennon, Martin and et al (1981). An evaluation of some conflation algorithms for information retrieval. *Journal of Information Science*, 3, pp.177-183.

Luhn, H. P. (1957). A statistical approach to the mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1 (4). pp. 309-317.

Matthews, P. H (1997). *Oxford concise dictionary of linguistics*. Oxford: Oxford University Press.

Montgomery, Christine A. (1972). Linguistics and information science. *Journal of the American Society for Information Science*, 23 (3), pp. 195-219.

Morfeq, Ali Hussein (1990). *Bayan: a text management system for Arabic engineering documents*. Ph.D. Thesis. Colorado: Colorado University.

Nasr, Raja T. (1967). *The structure of Arabic: from sound to sentence*. Beirut: Libraire De Liban.

Owens, Jonathan (1988). *The foundations of grammar: an introduction to medieval Arabic grammatical theory*. Amsterdam: John Benjamins Publishing Company.

Paice, Chris D. (1996). Method for evaluation of stemming algorithms based on error counting. *Journal of American Society for Information Science*, 47 (8), pp. 632-649.

Park, Taemin Kim (1993). The nature of relevance in information retrieval: an empirical study. *Library Quarterly*, 63 (3), pp.318-351.

Perez-Carballo, Jose and Tomek Strzalkowski (2000). Natural language information retrieval. progress report. *Information Processing and Management* 36 (2000), pp. 155-178

Plessis, B. Du. (1990). Producing Arabic document-dictionaries or concordance: prospects In: *Second Cambridge Conference Bilingual Computing in Arabic and English*

Popovic, Mirko and Peter Willett (1992). the effectiveness of stemming for natural-language access to Slovene textual data. *Journal of the American Society for Information Science*, 43 (5), pp.384-390.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14 (3), pp.130-137.

Robertson, S. E. and M. M. Hancock-Beaulieu (1992). In the evaluation of IR systems *Information Processing & Management*, 28 (4), pp. 457-466.

Robertson, Stephen E. (1981). The methodology of information retrieval Experiment. In: Karen Sparck Jones ed. *Information retrieval experiment*, London: Butterworth and Co.

Rowley, Jennifer (1998). *The electronic library*. London: Library Association Publishing.

Salton, Gerard. ed. (1971). *The SMART retrieval system: experiments in automatic document processing*, New Jersey: Prentice-Hall, INC.

Salton, Gerad and Michael J. McGill (1987). *Introduction to modern information retrieval* Singapore: McGraw-Hill International Book Co.

Saracevic, Tefko, (1975). Relevance: a review of and a framework for the thinking on the nation in information science. *Journal of American Society for Information Science*. 26 (6), pp. 321-343.

Saudi Soft (1995). Arabic full text retrieval. *Al Sharq Al Awsat Byte*, 1 (5), p.24.

Savoy, Jacqes (1993). Stemming of French words based on grammatical categories. *Journal of the American Society for Information Science*, 44 (1), pp.1-9.

Schamber, Linda (1994). Relevance and information behavior. *Annual Review of Information Science and Technologhy*, 29, pp.3-48.

Sakhr Software (1997). *Integrated Information Management system*. Cairo: Sakhr Software.

Smart, J. R (1986). *Arabic*. [Kent]: Hodder and Stonghton.

Smith, Peter (1996). *An introduction to knowledge engineering*. London: International Thomson Computer Press.

Sparck Jones, Karen and Martin Kay (1973). *Linguistics and information science*. New York: Academic Press.

Sparck Jones, Karen *ed.* (1981). *Information retrieval experiment*, London: Butterworth and Co.

Sparck Jones, Karen (2000). Further reflections on TREC. *Information Processing and Management*, 36 (2000), pp37-85.

Swanson, Don R. (1986). Subjective versus objective relevance in bibliographic retrieval systems. *Library Quarterly*, 56 (4), pp.389-398.

Tague, Jean M. (1981). The pragmatics of information retrieval experimentation. In: Karen Sparck Jones ed. *Information retrieval experiment*, London: Butterworth and Co.

Tague-Sutcliffe, Jean (1992). The pragmatic of information retrieval experiment, revisited *Information Processing & Management*, 28 (4), pp.464-490.

Tague-Sutcliffe, Jean (1996a). Information retrieval experimentation. In: Allen Kent and Carolyn M. Hall eds. *Encyclopedia of Library and Information Science*, vol. 57.

Tague-Sutcliffe, Jean M (1996b). Some perspectives on the evaluation on information retrieval systems. *Journal of American Society for Information Science*, 47 (1), pp 1-3.

Thalouth, Botrous and Abdullah Al Dannan (1989). A comprehensive Arabic morphological analyzer generator. In: P Mackay, ed. *Computers and the Arabic language*. New York: Hemisphere Publishing Corporation.

Tenopir, Carol and Jung Soon Ro (1990). *Full text databases*. New York: Greenwood Press.

Tenopir, Carol; Diane Nahl-Jakobovits and Dara Lee Howard (1990). Full text search strategies and modifications: the role of the searcher and the role of the system. In: *11th National Online Meeting*, pp.389-399.

Walker, Stephen and Richard M. Jones (1987). *Improving subject retrieval in online catalogues*. London: British Library Board.

Wall, A. E. (1989). *Morphology and grammar analysis program for the Arabic language*. MSc Thesis. Brunel: Brunel University.

Wallis, Peter, and James A. Thom (1996). Relevance judgments for assessing recall. *Information Processing & Management*, 32 (3), pp. 273-286.

Warner, Amy (1988). Linguistic theories for information retrieval. In: Patricia Hamalainen, Sininkka Koskiala and Aatoo J. Repo eds. *44th FID Conference and Congress*, Vol. 1

Warner, Amy J.; Ann Arbor and Aspen H. Wenzel (1991). A linguistic analysis and categorisation of nominal expressions. *ASIS'91*, pp. 186-191.

Wickens, G. M. (1980). *Arabic grammar*. Cambridge: Cambridge University Press.

Wright, W. (1974). *A grammar of the Arabic language*. Beirut: Libraire Du Liban.

Yang, Ming-Hsusan, Christopher C. Yang and Yi-Ming Chung (1997). A natural language processing based Internet agent. In *1997 IEEE*.

Yaqub, Imeal Badeea (1982).”"فقه اللغة العربية وخصائصها “ *Arabic philolgy and its characteristic*. Beirut: Dar eleilm llmalayeen.

Yaqub, Imeal Badeea (1993). “معجم الأوزان الصرفية” *Morphological forms dictionary*. Beirut: Alam Alkotub.

Appendices

Appendix 1

(Sample of the Arabic morphological forms representation as a list)

```
weight_3([L1,L2,L3],[L1,L2,L3]).

weight_4([L1,L2,'ا',L3],[L1,L2,L3]).
weight_4([L1,L2,'ي',L3],[L1,L2,L3]).
weight_4([L1,L2,'و',L3],[L1,L2,L3]).
weight_4(['م',L1,L2,L3],[L1,L2,L3]).
weight_4(['أ',L1,L2,L3],[L1,L2,L3]).
weight_4([L1,L2,L3,'ة'],[L1,L2,L3]).
weight_4([L1,L2,L3,'ه'],[L1,L2,L3]).
weight_4([L1,L2,L3,'ى'],[L1,L2,L3]).
weight_4([L1,L2,L3,'ل'],[L1,L2,L3]).
weight_4(['ت',L1,L2,L3],[L1,L2,L3]).
weight_4([L1,'ا',L2,L3],[L1,L2,L3]).
weight_4([L1,'ي',L2,L3],[L1,L2,L3]).
weight_4([L1,L2,L3,'ل'],[L1,L2,L3]).

weight_5([L1,L2,'ا',L3,'ة'],[L1,L2,L3]).
weight_5([L1,L2,'ا',L3,'ه'],[L1,L2,L3]).
weight_5([L1,'ا','ي',L3,'ه'],[L1,'و',L3]).
weight_5([L1,L2,'ا',L3,L4],[L1,L2,L3,L4]).
weight_5(['ت',L1,L2,L3,'ة'],[L1,L2,L3]).
weight_5(['م',L1,'ا',L2,'ة'],[L1,'ا',L2]).
weight_5(['ت',L1,'ا',L2,L3],[L1,L2,L3]).
weight_5([L1,L2,'و',L3,'ة'],[L1,L2,L3]).
weight_5([L1,L2,L3,'ء','ا'],[L1,L2,L3]).
weight_5(['أ',L1,L2,'ا',L3],[L1,L2,L3]).
weight_5([L1,L2,L3,'ت','و'],[L1,L2,L3]).
weight_5([L1,L2,'و',L3,'ة'],[L1,L2,L3]).
weight_5(['م',L1,L2,L3,'ة'],[L1,L2,L3]).
weight_5([L1,L2,'ي',L3,'ة'],[L1,L2,L3]).
weight_5([L1,L2,'ي',L3,'ه'],[L1,L2,L3]).
weight_5([L1,'ا',L2,L3,'ة'],[L1,L2,L3]).
weight_5([L1,L2,L3,'ة','ل'],[L1,L2,L3]).
weight_5([L1,'و',L2,L3,'ة'],[L1,L2,L3]).
weight_5([L1,'ي',L2,L3,'ة'],[L1,L2,L3]).
weight_5(['ا',L1,L2,'ا',L3],[L1,L2,L3]).
weight_5(['إ',L1,L2,'ا',L3],[L1,L2,L3]).
weight_5(['ت',L1,L2,'ا',L3],[L1,L2,L3]).
weight_5([L1,L2,L3,'ل','ا'],[L1,L2,L3]).
```

```
weight_5([L1,'ي',L2,'ا',L3],[L1,L2,L3]).
weight_5([L1,L2,L3,'ن','ا'],[L1,L2,L3]).
weight_5([L1,L2,L3,'ت','ا'],[L1,L2,L3]).
weight_5(['ت',L1,L2,'ي',L3],[L1,L2,L3]).
weight_5(['ت',L1,L2,L3,'ل'],[L1,L2,L3]).
weight_5(['ت',L1,L2,L3,'ي'],[L1,L2,L3]).
weight_5(['ت',L1,L2,L3,'ت'],[L1,L2,L3]).
weight_5(['ت',L1,L2,'ي',L3],[L1,L2,L3]).
weight_5(['ت',L1,'و',L2,L3],[L1,L2,L3]).
weight_5(['م','ت',L1,L2,L3],[L1,L2,L3]).
weight_5(['ت',L1,'ن',L2,L3],[L1,L2,L3]).
weight_5(['م',L1,'ي',L3,'ة'],[L1,'ي',L3]).
weight_5(['م',L1,L2,'و',L3],[L1,L2,L3]).
weight_5(['م',L1,L2,'ي',L3],[L1,L2,L3]).
weight_5(['م',L1,L2,'ا',L3],[L1,L2,L3]).
weight_5([L1,'ا',L2,'و',L3],[L1,L2,L3]).
weight_5([L1,L2,L3,'ي','ل'],[L1,L2,L3]).
weight_5([L1,L2,L3,'ل','ي'],[L1,L2,L3]).
weight_5([L1,L2,L3,'ة','ي'],[L1,L2,L3]).

weight_6(['م',L1,'ت',L2,L3,'ة'],[L1,L2,L3]).
weight_6(['م',L1,'ت',L2,L3,'ه'],[L1,L2,L3]).
weight_6([L1,L2,'ا',L3,'ة','ي'],[L1,L2,L3]).
weight_6(['ا',L1,'ت',L2,'ا',L3],[L1,L2,L3]).
weight_6(['م',L1,'ا',L2,'ي',L3],[L1,L2,L3]).
weight_6(['ت',L1,'ا',L2,'ي',L3],[L1,L2,L3]).
weight_6(['م',L1,'ا',L2,L3,'ة'],[L1,L2,L3]).
weight_6(['ن','ا',L1,L2,'ا',L3],[L1,L2,L3]).
weight_6(['ا',L1,L2,'ا','و',L3],[L1,L2,L3]).
weight_6(['ا',L1,L2,'ا','م',L3],[L1,L2,L3]).
weight_6(['ت',L1,L2,'ا',L3,'ة'],[L1,L2,L3]).
weight_6([L1,L2,L3,'ء','ا','ي'],[L1,L2,L3]).
weight_6(['ا',L1,L2,'ا','و',L3],[L1,L2,L3]).
weight_6([L1,L2,L3,'ل','ي','ل'],[L1,L2,L3]).
weight_6([L1,L2,'و',L3,'ة','ي'],[L1,L2,L3]).
weight_6([L1,L2,'ي',L3,'ء','ا'],[L1,L2,L3]).
weight_6([L1,L2,L3,'ة','ل','و'],[L1,L2,L3]).
weight_6([L1,L2,'ي',L3,'ة','ي'],[L1,L2,L3]).
weight_6([L1,L2,L3,'ة','ي','ن'],[L1,L2,L3]).
weight_6(['ا',L1,L2,'ا','و',L3],[L1,L2,L3]).
weight_6([L1,L2,L,L3,'ت','ا'],[L1,L2,L3]).

weight_7(['ت','س','ا',L1,L2,L3,'ة'],[L1,L2,L3]).
weight_7(['ت','س','ا',L1,L2,'ء','ا'],[L1,L2,'ي']).
weight_7(['ت','س','ا',L1,L2,'ء','ا'],[L1,L2,'ا']).
```

```
weight_7(['ت','س','ا',L1,L2,'ا',L3],[L1,L2,L3]).
weight_7(['ت','س','إ',L1,L2,'ا',L3],[L1,L2,L3]).
weight_7(['ا',L1,L2,'ي',L3,'ل','ا'],[L1,L2,L3]).
weight_7(['ا',L1,L2,'ت',L3,'ء','ا'],[L1,L2,L3]).
weight_7(['ا',L1,L2,'ا','ع','ي',L3],[L1,L2,L3]).
weight_7(['ا',L1,'م',L2,L3,'ل','ا'],[L1,L2,L3]).
weight_7(['ا',L1,L2,'ن',L3,'ل','ا'],[L1,L2,L3]).
weight_7(['ا',L1,L2,'ا','م','ن',L3],[L1,L2,L3]).
weight_7(['ا',L1,L2,'و',L3,'ل','ا'],[L1,L2,L3]).
weight_7(['ا',L1,L2,'ي',L3,'ل','ا'],[L1,L2,L3]).
weight_7(['ن','ا',L1,L2,'ا',L3,'ة'],[L1,L2,L3]).
weight_7(['ا',L1,L2,'ا','ع','ي',L3],[L1,L2,L3]).
weight_7(['ا',L1,L2,'ا','ع','ي',L3],[L1,L2,L3]).
weight_7(['إ',L1,L2,'ن',L3,'ء','ا'],[L1,L2,L3]).
weight_7(['ا',L1,L2,'ي',L3,'ء','ا'],[L1,L2,L3]).
weight_7(['م',L1,L2,'و',L3,'ة','ي'],[L1,L2,L3]).

weight_8(['ا',L1,'ت',L2,'ا',L3,'ة','ي'],[L1,L2,L3]).
```

Appendix 2

(Sample of Arabic text and Prolog database representation)

```
record(
number(270),
author('مارميون ، دان '),
title('توصيل محطات الفهرس العام شبكيّا'),
journal('المكتبات في الحاسوبات'),
volume('مج١٧، ع٥ ( مايو ١٩٩٧م )'),
pages('ص ٢٦ - ٢٩'),
abstract(['
في هذه المقالة يتحدث الكاتب عن تجربة قامت بها جامعة ميتشيجان الأمريكية لتغيير نظـــام حاســوبات
محطات العمل الخاصة بالفهارس الرقمية التي يمكن استخدامها عن طريق الاتصال المباشر، فلقد قـــامت
الجامعة بتغيير محطاتها القديمة بحاسوبات شخصية لها برامجها الخاصة .وللتغلب على مشكلة الحاجة إلى
برمجة كل حاسوب شخصي على حدة وكذلك الحاجة إلى الصيانة الفردية لكل حاسوب جديد سواء كـــانت
الصيانة تخص المكونات المادية للحاسوب أو برامجه، لقد قامت الجامعة بتوصيل تلك الحاســوبات كلـــها
بشبكة و حدة ووضع كل البرامج الخاصة بالفهرس على الحاسوب المركزي لتلك الشبكة .ويبدأ الكاتب مقاله
بالقول إن المكتبات الكبرى خاصة الأكاديمية منها يمكن أن تمتلك مئات من محطات العمل العامة للتعامل
مع فهرس المكتبة .وفي الأيام الماضية كانت محطات العمل العامة تلك مجرد طرفيات غير ذكية ) بدون
ذاكرة خاصة بها ( وكانت كلها تتصل بحاسوب واحد مركزي لكن مكتبة المستقبل ) وفي الحقيقة كثـــير
من مكتبات الحاضر   يتحتم عليها أن تستبدل تلك الطرفيات الغبية بالحاسوبات الصغيرة ) الشخصية. (
ذلك ؛ فإن ما كان لا يحتاج إلى صيانة تقريباً يصبح فجأة بيئة تحتاج إلى عمالة مكثفة .ولابد أن توضع
البرامج في تلك الحاسوبات حتى يمكن لها أن تعمل كمحطات عامة للتعامل مع الفهرس ) الفهرس العـــام
ويعني تركيب تجديدات لتلك البرامج تكرار نفسها العمليات مع كل حاسوب؛ فإذا فشل القرص الصلب: .
فإن عليك أن تستبدل معه كل البرامج مرة أخرى ، ولسوف ترى أنه لابد من طريقة أفضل. والحقيقة هي
أنه توجد طريقة أفضل وهي عملية شبكة محطات الفهارس العامة المتاحة للجمهور معاً .وهذا ما قامت به
جامعة غرب ميتشيجان وما يتناوله المقال بالتفصيل
']) .
```

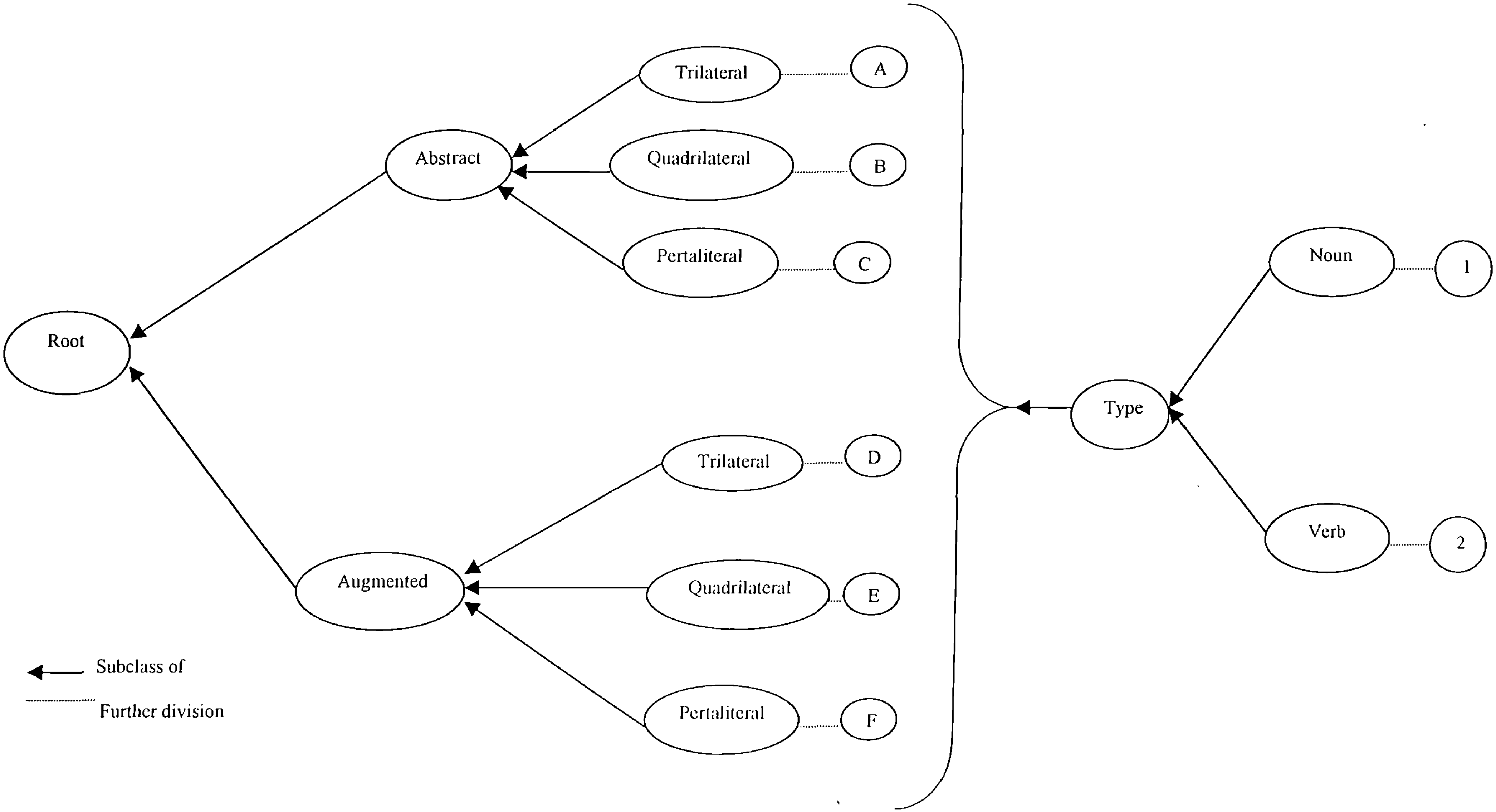
Appendix 3

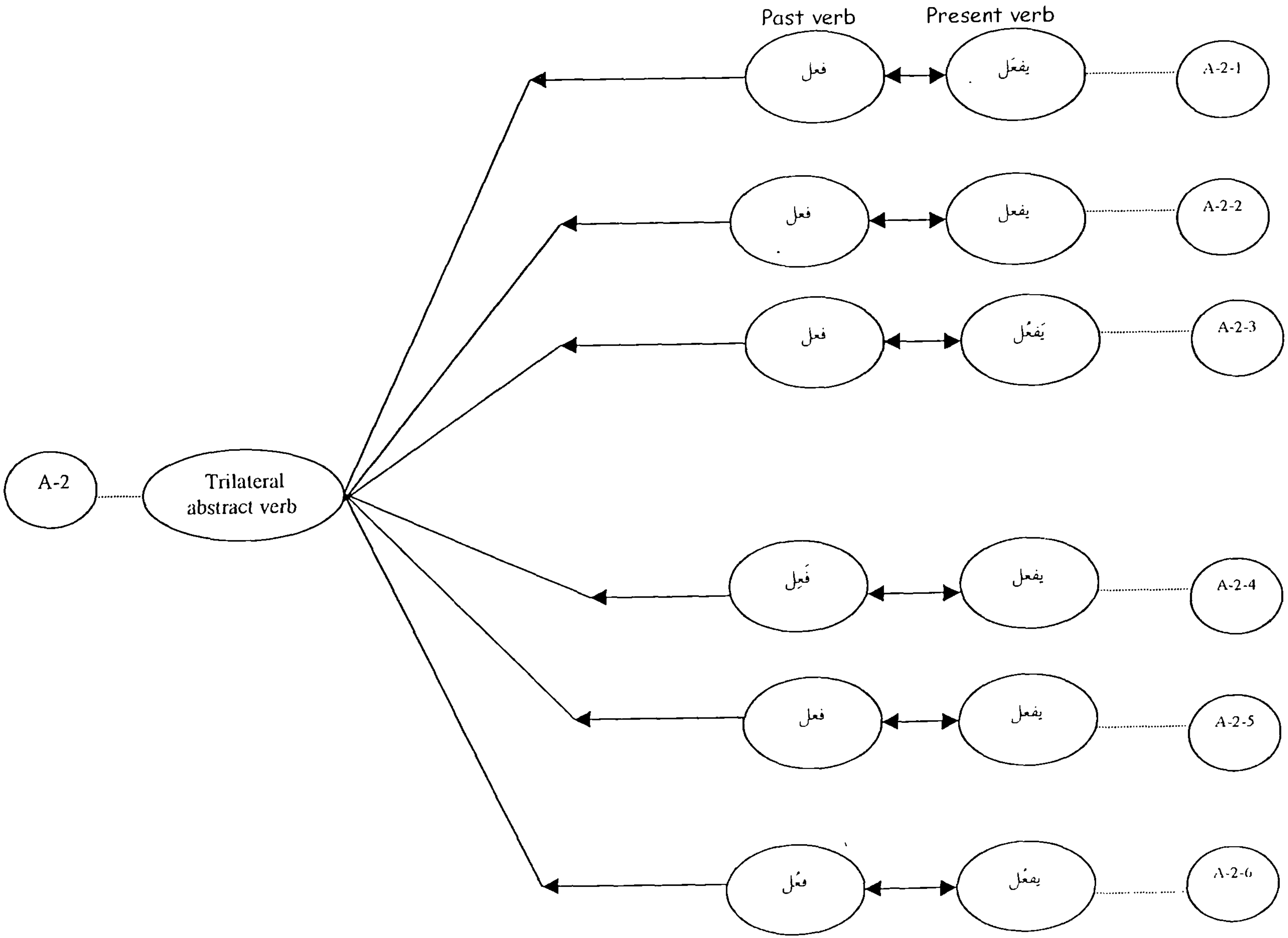
(Arabic query statements and their English translation)

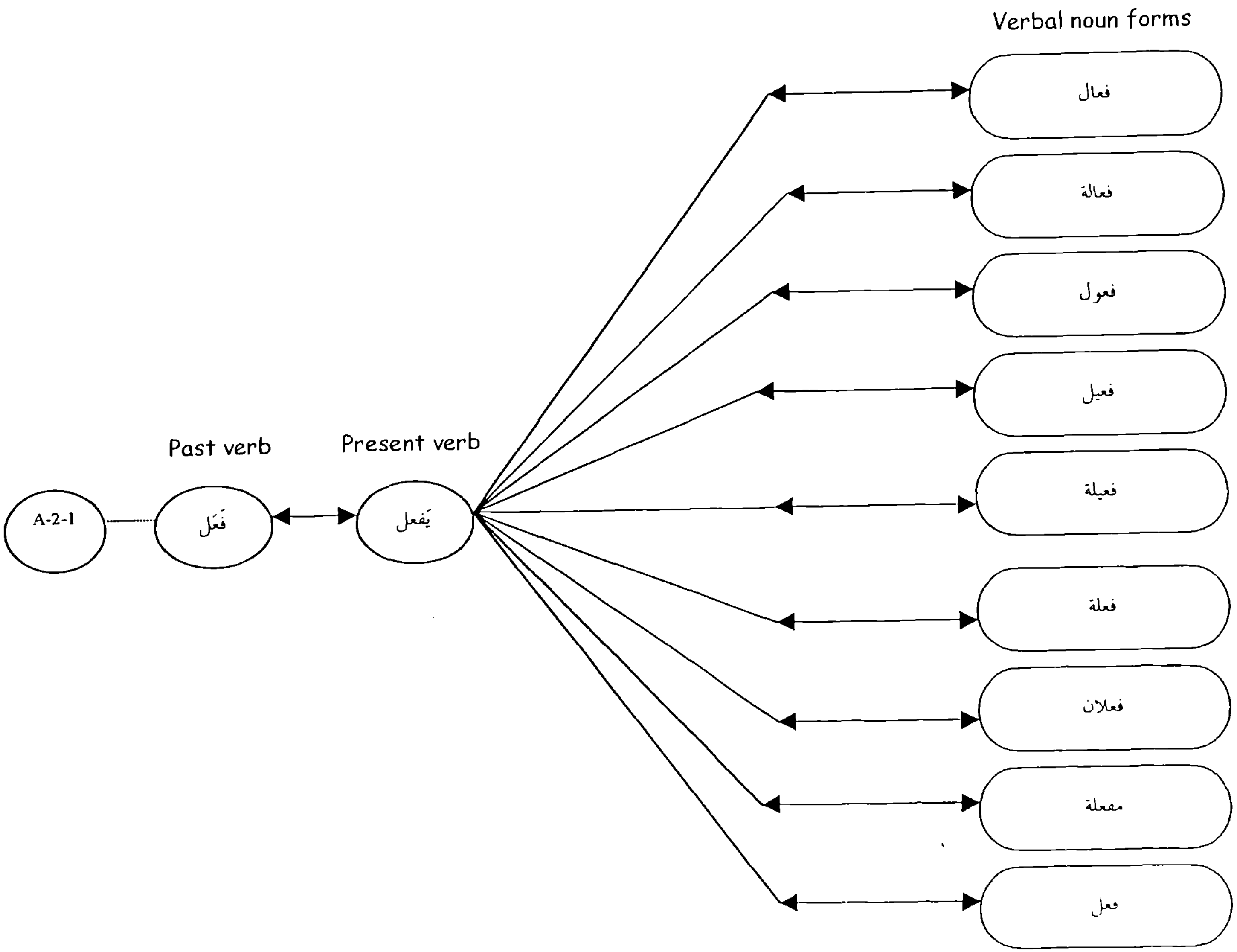
No.	English translation	Arabic queries
1	Drugs	المخدرات
2	Plants	النباتات
3	Medicine	الطب
4	Domestic animals	الدواجن
5	Milk	الألبان
6	Agrology	التربة
7	Crops	المحاصيل
8	Drink water	مياه الشرب
9	Radiation	الإشعاع
10	Waste	النفايات
11	Air pollution	تلوث الهواء
12	Environment protection	حماية البيئة
13	Job market	سوق العمل
14	Engineering education	التعليم الهندسي
15	University lecturer	الأستاذ الجامعي
16	Cataloging	الفهرسة
17	Electronic journals	المجلات الإلكترونية
18	Copyright Deposit	الإيداع
19	University libraries	المكتبات الجامعية
20	Author's rights	حقوق المؤلف
21	Electronic library	المكتبة الإلكترونية
22	Human's rights	حقوق الإنسان
23	Books	الكتب الممنوعة
24	Almojtam journal	مجلة المجتمع
25	Currency	العملات
26	Banks	المصارف
27	Installment sale	التقسيط
28	Interest	الربا
29	Loans	الديون
30	Tax	الضرائب
31	Contracts	العقود
32	Money	النقود

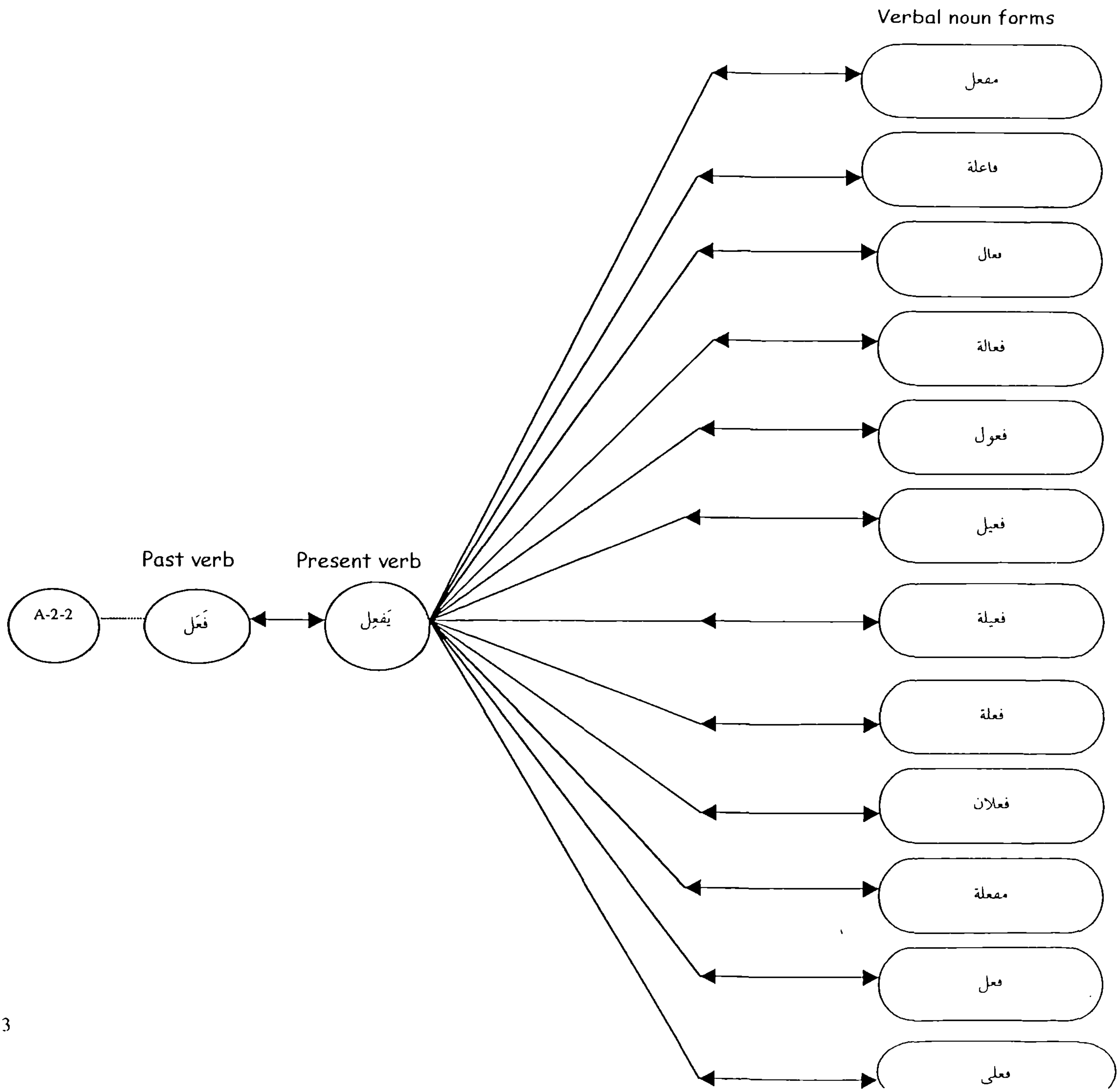
Appendix 4

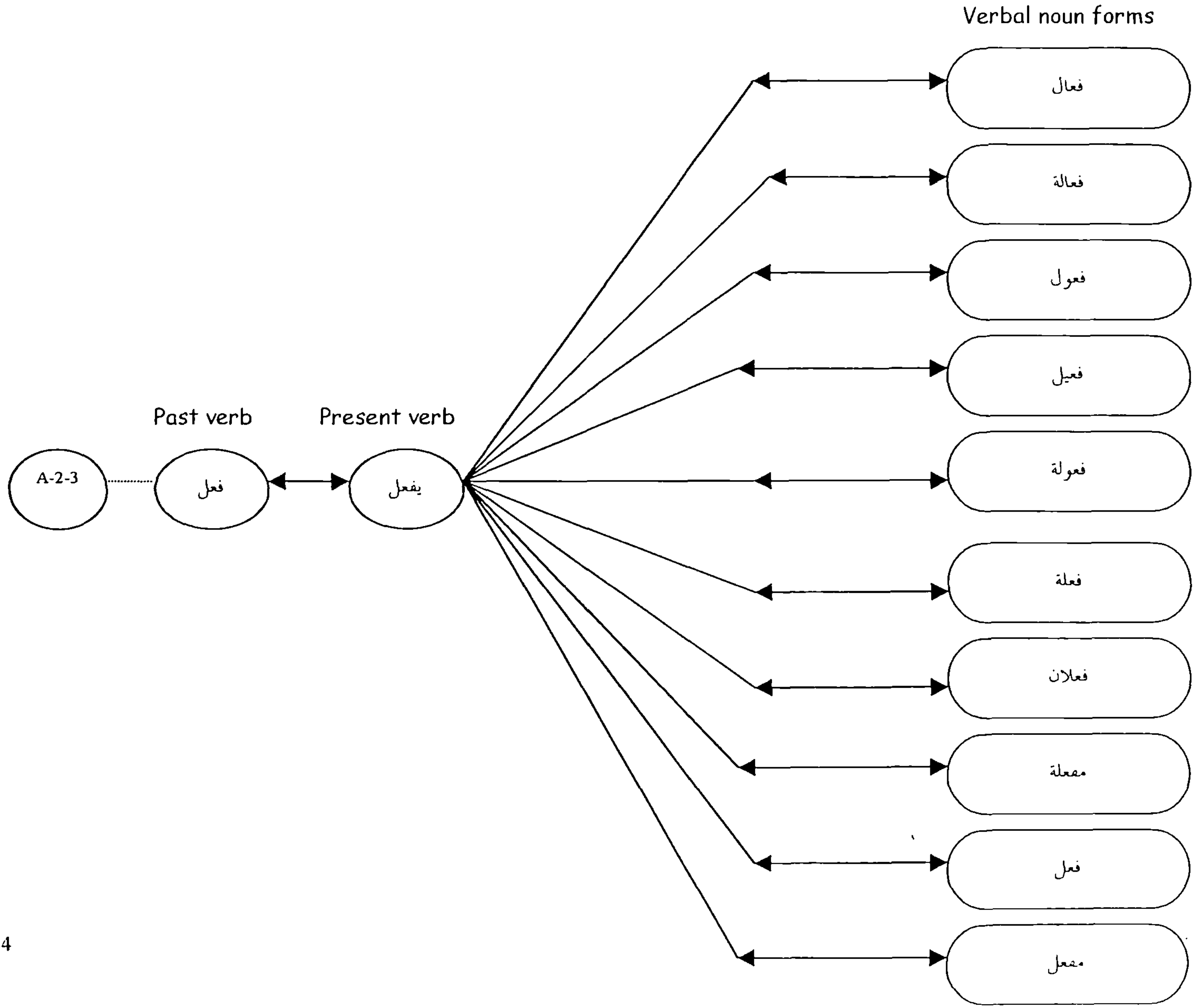
Examples of derivational path links of morphological forms

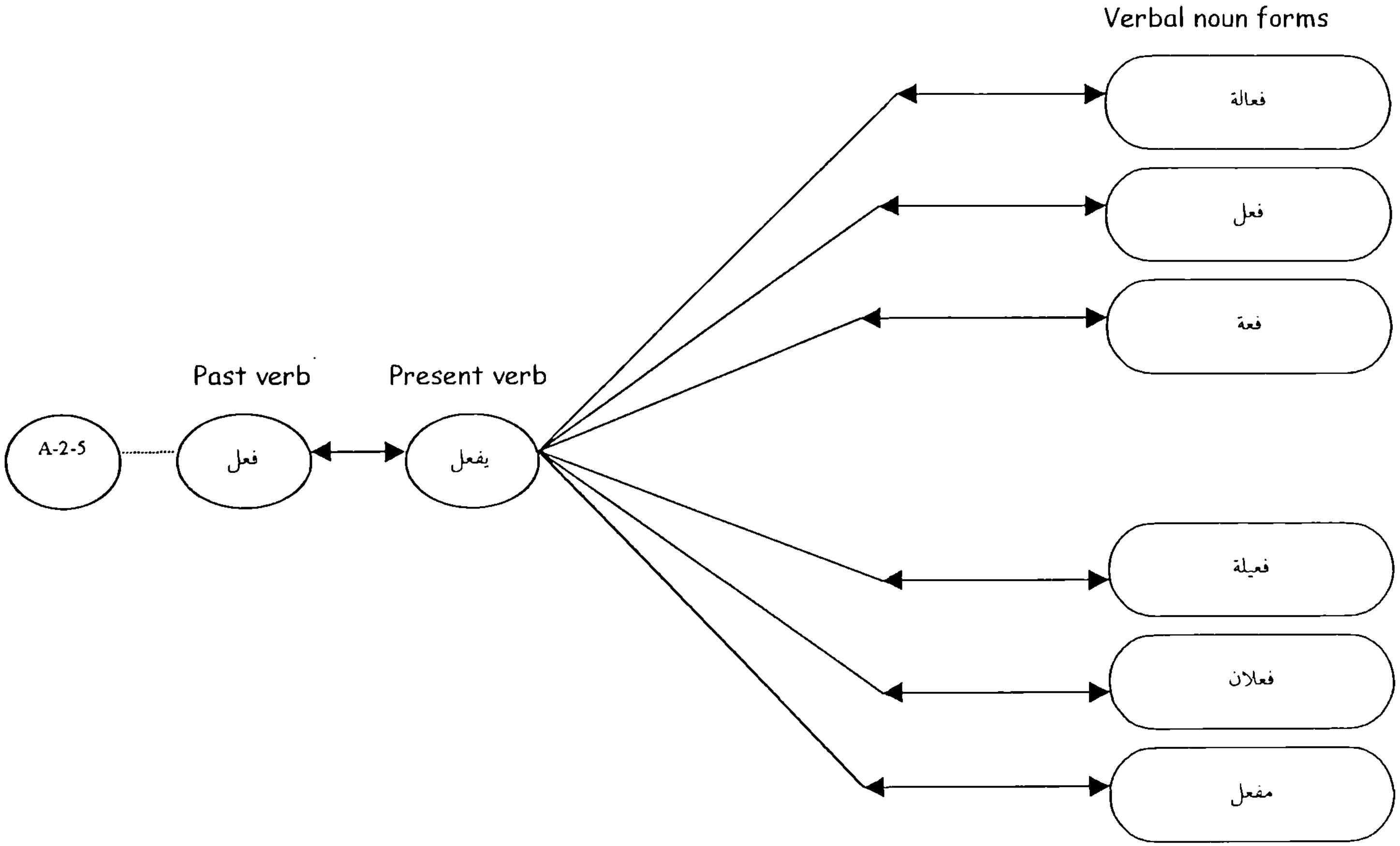

Root
Abstract
Augmented
Trilateral
Quadrilateral
Pertaliteral
A
B
C
Trilateral
Quadrilateral
Pertaliteral
D
E
F
Type
Noun
Verb
1
2
Subclass of
Further division

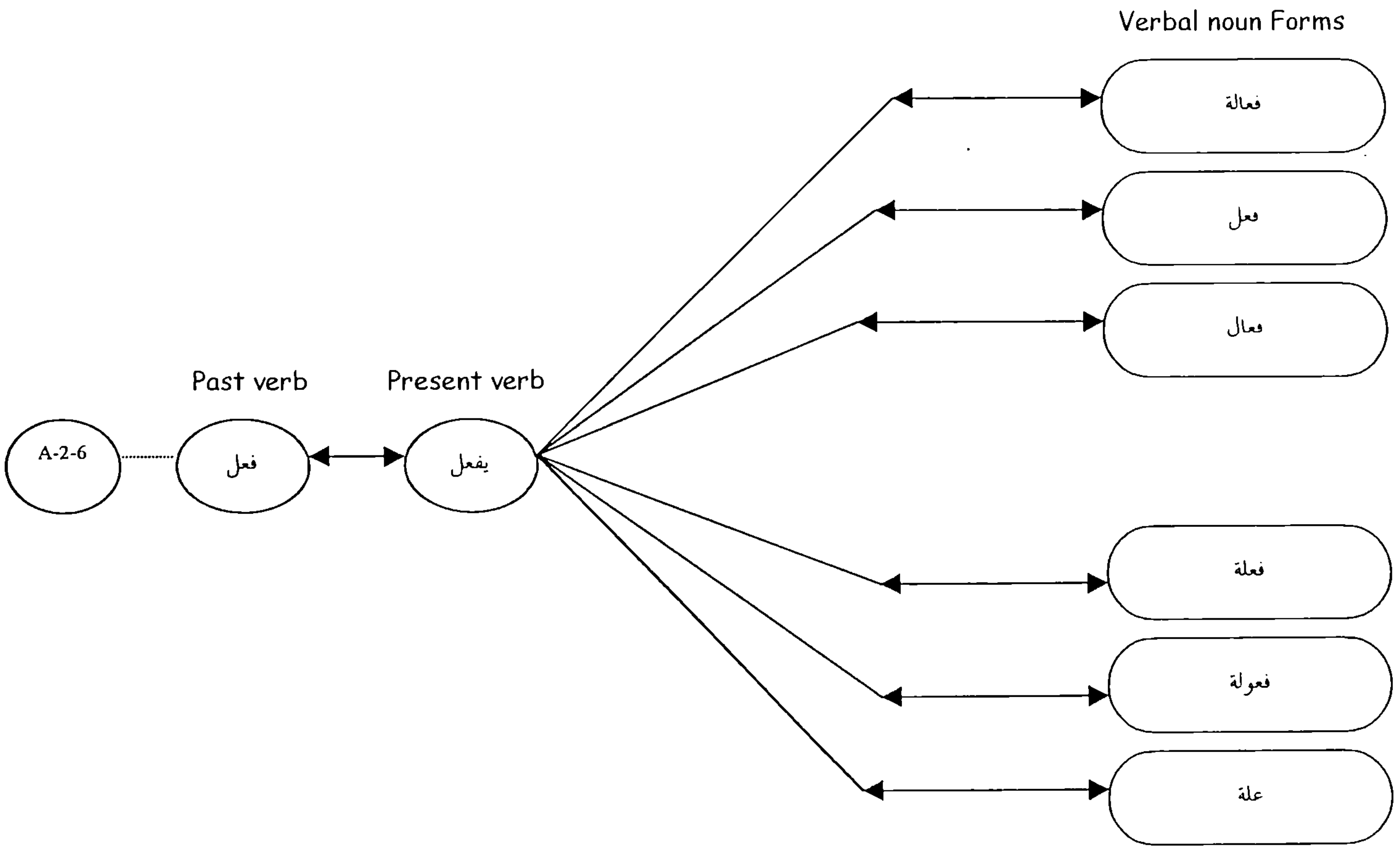

Past verb
Present verb
A-2
Trilateral abstract verb
فعل
يفعَل
A-2-1
فعل
يفعل
A-2-2
فعل
يَفعُل
A-2-3
فَعِل
يفعل
A-2-4
فعل
يفعل
A-2-5
فعُل
يفعُل
A-2-6

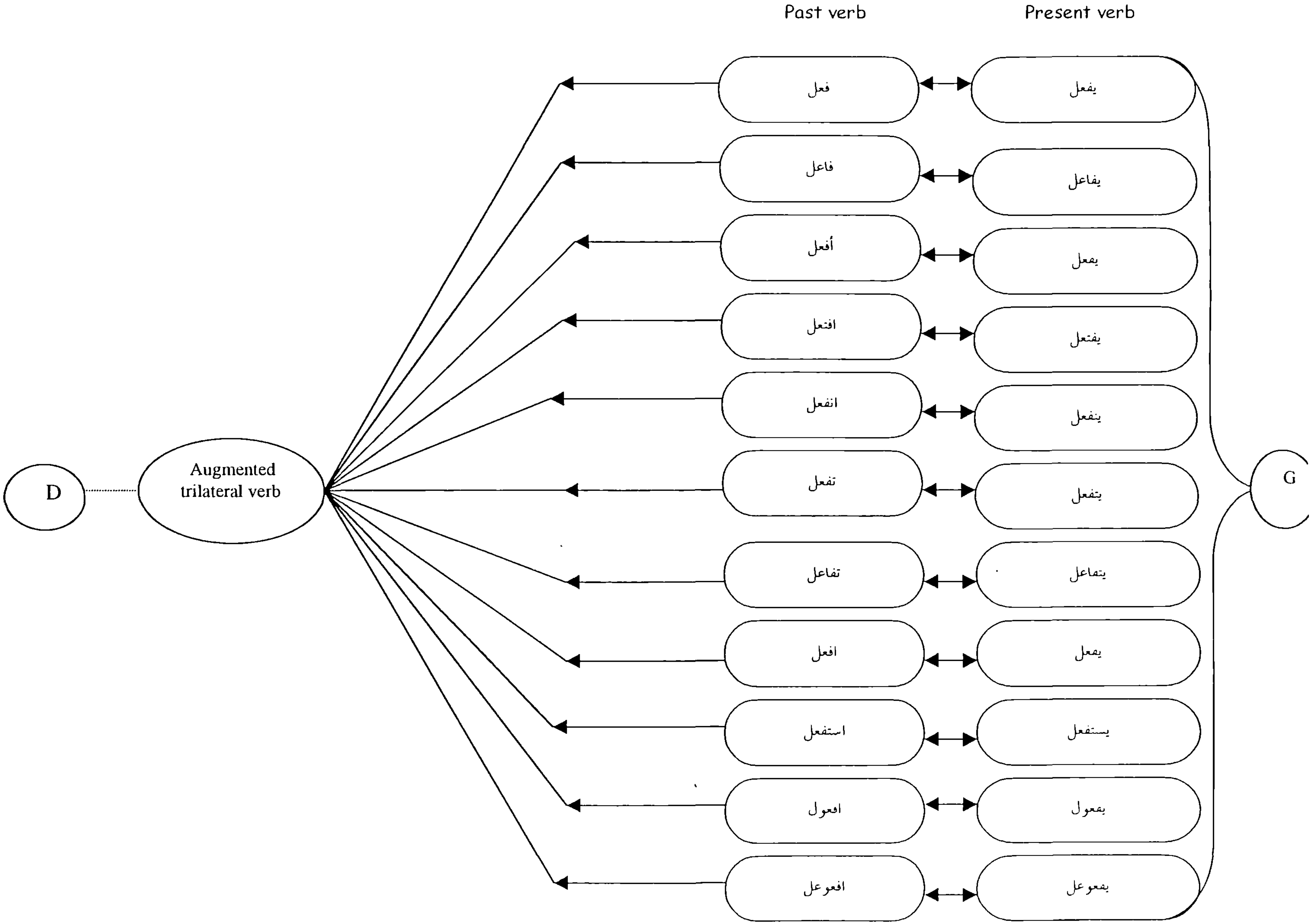

Verbal noun forms
Past verb
Present verb
A-2-1
فَعَل
يَفعل
فعال
فعالة
فعول
فعيل
فعيلة
فعلة
فعلان
مفعلة
فعل

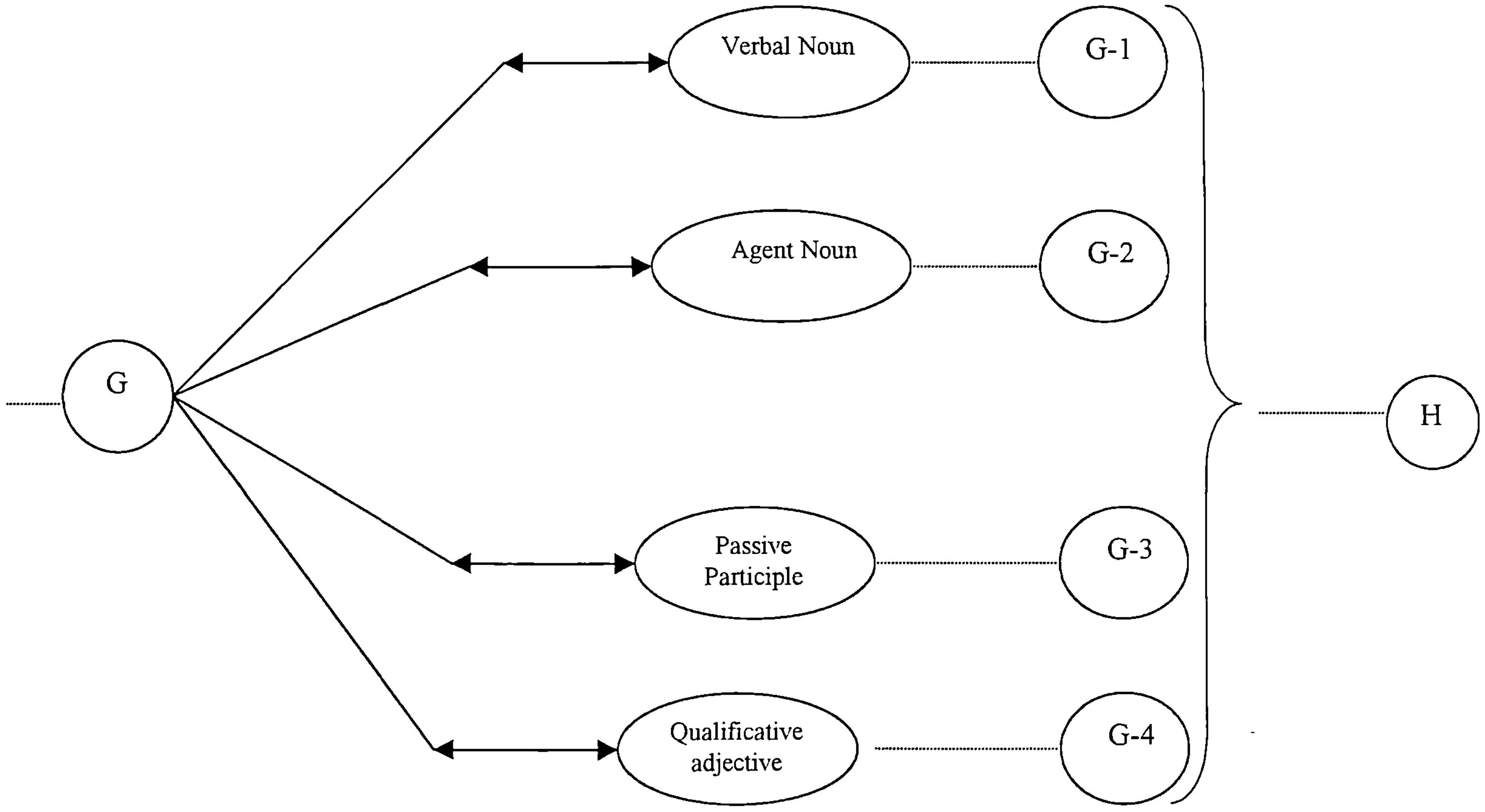

Verbal noun forms
A-2-2
Past verb
فَعَل
Present verb
يَفعِل
مفعل
فاعلة
فعال
فعالة
فعول
فعيل
فعيلة
فعلة
فعلان
مفعلة
فعل
فعلى

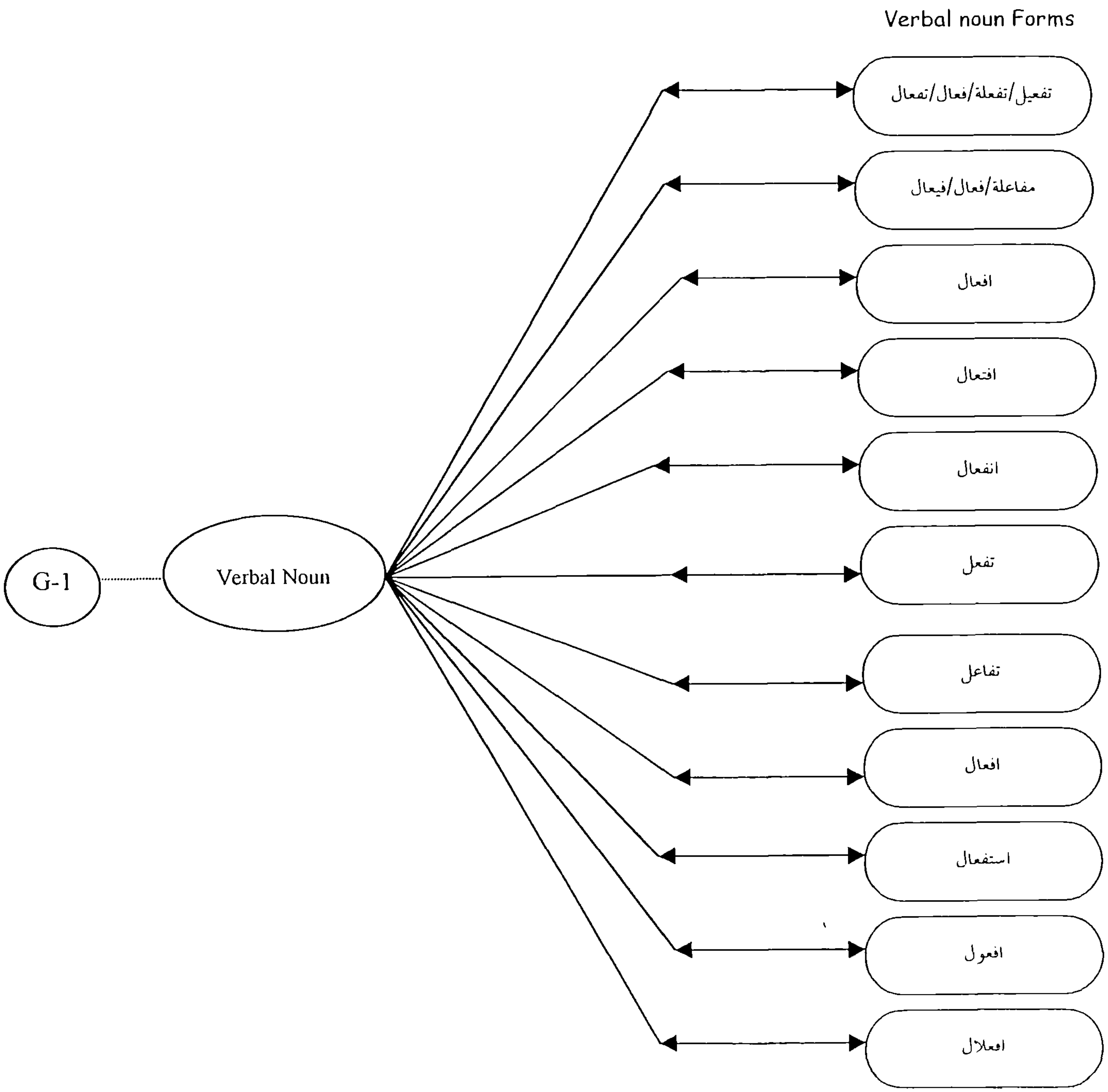

Verbal noun forms
Past verb
Present verb
A-2-3
فعل
يفعل
فعال
فعالة
فعول
فعيل
فعولة
فعلة
فعلان
مفعلة
فعل
مفعل

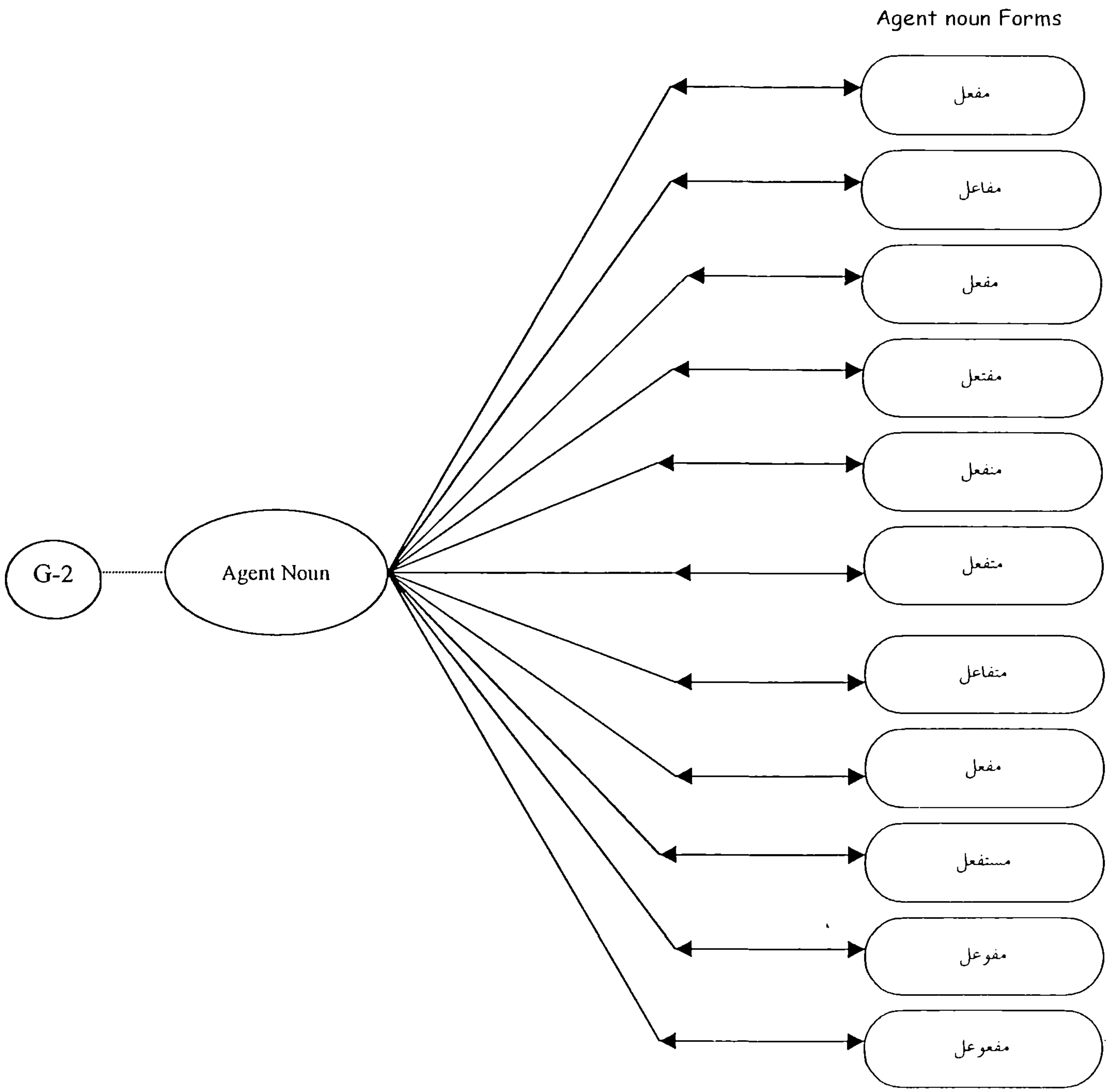

Verbal noun forms
Past verb
Present verb
A-2-5
فعل
يفعل
فعالة
فعل
فعة
فعيلة
فعلان
مفعل


Verbal noun Forms
Past verb
Present verb
A-2-6
فعل
يفعل
فعالة
فعل
فعال
فعلة
فعولة
علة

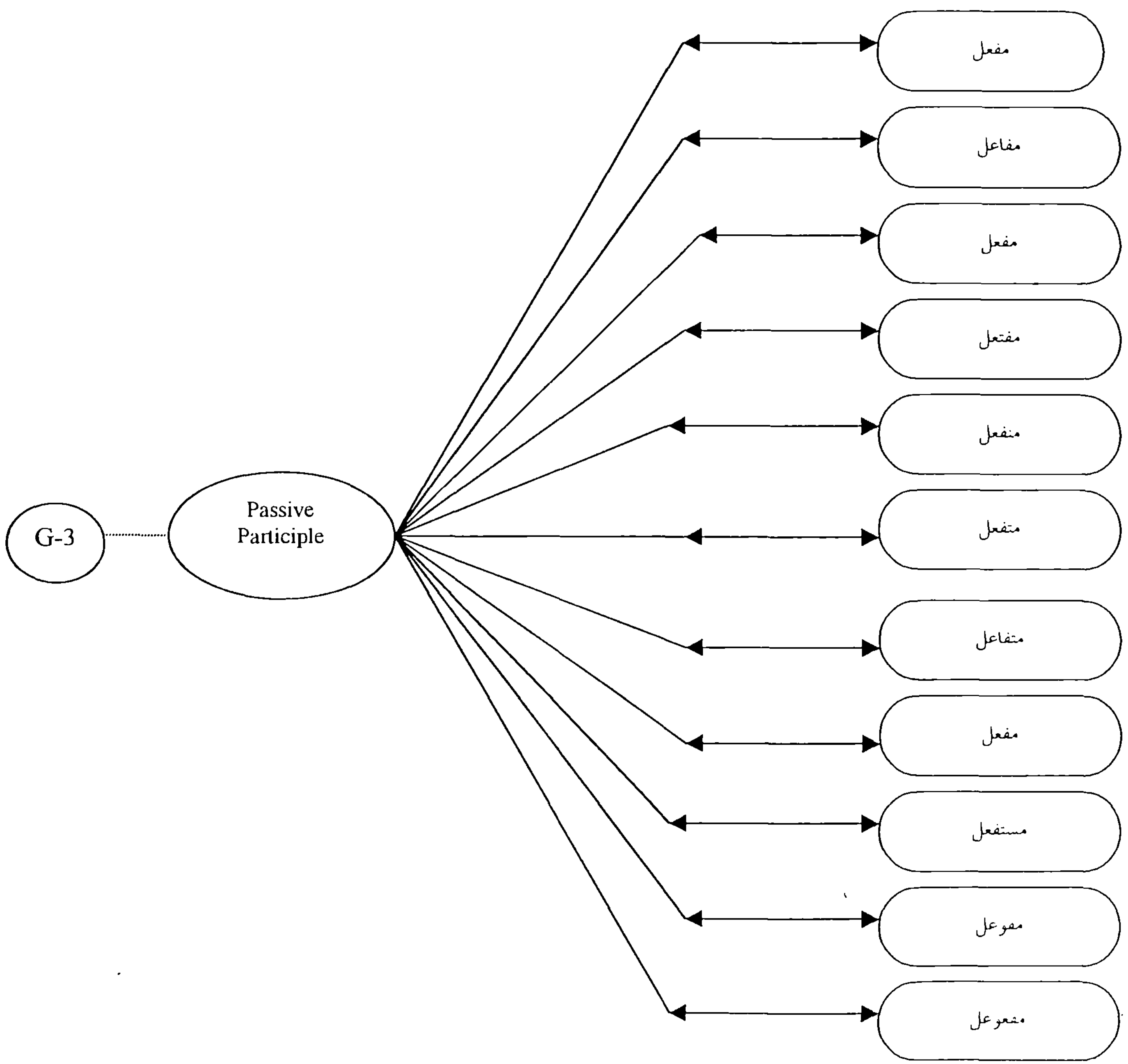

Past verb
Present verb
D
Augmented trilateral verb
فعل
يفعل
فاعل
يفاعل
أفعل
يفعل
افتعل
يفتعل
انفعل
ينفعل
تفعل
يتفعل
تفاعل
يتفاعل
افعل
يفعل
استفعل
يستفعل
افعول
يفعول
افعوعل
يفعوعل
G


Verbal Noun
G-1
Agent Noun
G-2
G
H
Passive Participle
G-3
Qualificative adjective
G-4

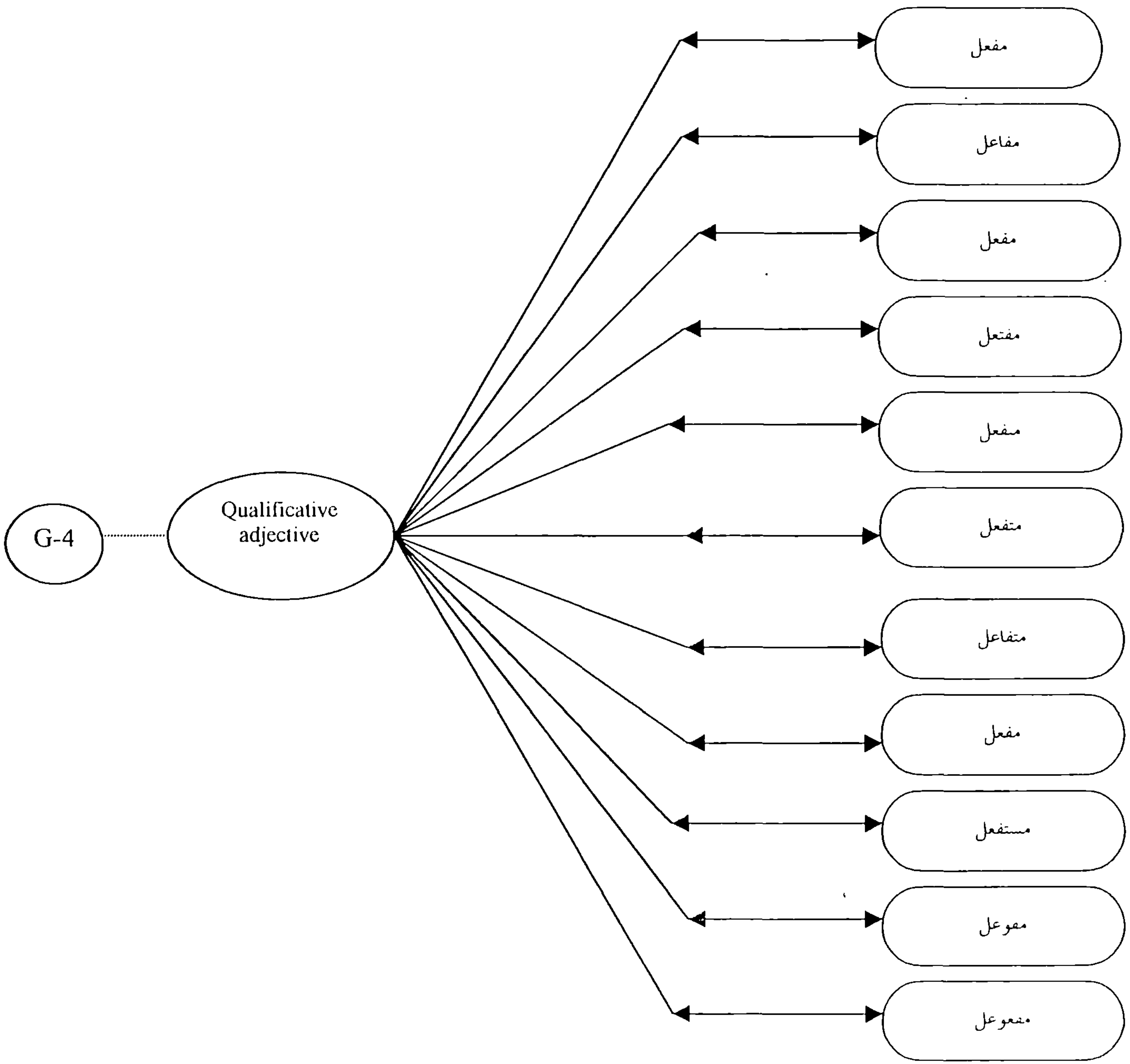

Verbal noun Forms
G-1
Verbal Noun
تفعيل/تفعلة/فعال/تمعال
مفاعلة/فعال/فيعال
افعال
افتعال
انفعال
تفعل
تفاعل
افعال
استفعال
افعول
افعلال

Agent noun Forms
G-2
Agent Noun
مفعل
مفاعل
مفعل
مفتعل
منفعل
متفعل
متفاعل
مفعل
مستفعل
مفوعل
مفعوعل

Passive participle noun Forms



Qualificative adjective Forms


G-4
Qualificative adjective
مفعل
مفاعل
مفعل
مفتعل
مفعل
متفعل
متفاعل
مفعل
مستفعل
مفوعل
متفوعل